Advances in the sociology of trust and cooperation: theory, experiments, and field studies

Buskens, Vincent (Ed.); Corten, Rense (Ed.); Snijders, Chris (Ed.)

Veröffentlichungsversion / Published Version
Sammelwerk / collection

Empfohlene Zitierung / Suggested Citation:
Buskens, V., Corten, R., & Snijders, C. (Eds.). (2020). *Advances in the sociology of trust and cooperation: theory, experiments, and field studies*. Berlin: De Gruyter. https://doi.org/10.1515/9783110647495

Advances in the Sociology of Trust and Cooperation

Advances in the Sociology of Trust and Cooperation

Theory, experiments, and field studies

Edited by
Vincent Buskens, Rense Corten and Chris Snijders

DE GRUYTER

ISBN 978-3-11-064701-3
e-ISBN (PDF) 978-3-11-064749-5
e-ISBN (EPUB) 978-3-11-064761-7
DOI https://doi.org/10.1515/9783110647495

Library of Congress Control Number: 2020946148

Bibliographic information published by the Deutsche Nationalbibliothek
The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data are available on the Internet at http://dnb.dnb.de.

Cover image: Viesinsh/iStock/Getty Images Plus
Typesettung: Integra Software Services Pvt. Ltd.
Printing and Binding: CPI books GmbH, Leck

www.degruyter.com

Contents

Part III: **Field Studies**

Vincent Buskens, Rense Corten and Chris Snijders

1 Complementary Studies on Trust and Cooperation in Social Settings: An Introduction

> Rigorous sociologists should develop sound theoretical predictions to be tested with high-quality empirical research rather than produce 'teutonischer Tiefsinn' devoid of empirical content.
>
> Werner Raub

1.1 Background

The issue of cooperation has been a core topic in the social sciences for a long time, and for good reasons. Many societal matters share that everyone involved knows what they would prefer to see happening, but the incentives are such that this is hard or even impossible to achieve. Examples are abundant, and play at different levels of granularity. At the societal level, one could think of (trying to prevent) the depletion of collective resources or transitioning to a more sustainable society. At the level of organizations, one could think of trying to overcome the impulse to benefit in a business relation at the expense of the other party, or of the tendency for businesses to use legal constructs to evade taxes. Regardless of the grandiosity of these cooperation problems, similar arguments play a role in the provision of collective goods in neighborhoods or households, and even within a single person there can be friction between the current and the future self.

The academic literature has coined the term "social dilemmas" for these kinds of interactions in which sensible decisions by individual actors lead to an outcome that is inferior for all (see Raub, Corten, and Buskens 2015 for a more technical definition). For those unacquainted with the general topic, this sounds hard to imagine. How can it be that we end up in a situation where everybody agrees that an alternative outcome was possible that is better for everybody? Nevertheless, many well-known cooperation problems share this feature. The canonical example of such a problem is the famous Prisoner's Dilemma (Luce and Raiffa 1957: 94–95; Axelrod 1984). In this game for two actors, the two actors simultaneously choose between two actions, *Cooperate* or *Defect*, and the four outcomes that can result from these choices have benefits for both actors as in the left panel of Figure 1.1. Both actors, considering the potential choices of the other player, will conclude that to defect benefits them more than to cooperate, irrespective of what the other actor does,

Vincent Buskens, Rense Corten, Department of Sociology / ICS, Utrecht University
Chris Snijders, Human Technology Interaction, Eindhoven University of Technology

https://doi.org/10.1515/9783110647495-001

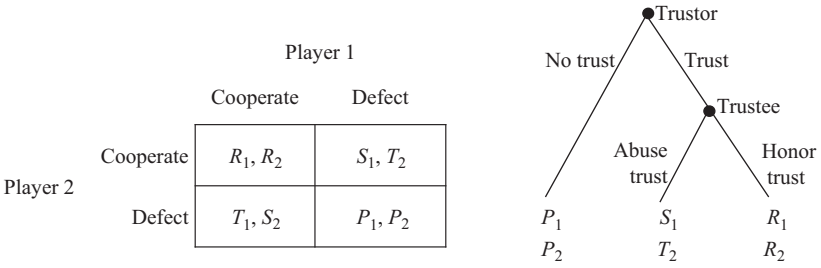


Figure 1.1: Prisoner's Dilemma (left) and Trust Game (right) ($S_i < P_i < R_i < T_i$, $i = 1,2$).

and hence to defect is the rational course of action for each of them. Two actors who think this way will end up with the payoffs that go with mutual defection, even though mutual cooperation would have yielded a better outcome to both.

Another type of exchange that differs from the Prisoner's Dilemma in important aspects but nevertheless shares its social dilemma nature is the *Trust Game* (Dasgupta 1988), illustrated in the right panel of Figure 1.1. In this game, the moves are sequential rather than simultaneous. In the first move, the "trustor" chooses whether or not to place trust in the "trustee." If the trustor indeed decides to place trust, the trustee decides whether to honor or abuse it. The outcomes of the Trust Game are such that after trust is placed, abusing trust is more beneficial for the trustee than honoring trust. The trustor, anticipating that the trustee will abuse trust if it is placed, will reasonably decide not to trust the trustee. As in the Prisoner's Dilemma, both actors end up in a situation that is suboptimal as mutual cooperation (the trustor placing trust and the trustee honoring trust) would be a better outcome for both.

The explanation of these, already very well-known, abstract interactions immediately suggests both how theoretical arguments about cooperation problems can take shape and how scholars have tried to tackle the analysis of cooperation and trust problems. For one, these games make clear that not all social dilemmas are equal. Some fit better with the Prisoner's Dilemma, where choice is simultaneous, whereas others resemble a Trust Game, where choice is sequential. Other types of games are often natural extensions of these archetypical ones. It can be sensible to assume more than two actors, more than just two behavioral choices, different kinds of payoffs, uncertainty about choices or payoffs or the type of other players, to assume repetition of interactions, embeddedness of the interactions in a larger setting, or different rules about behavior or outcomes.

Theoretical social scientists have covered a variety of such models, trying to identify the conditions under which cooperation may emerge. Such explanations may take place at several levels (cf. Kollock 1998). At an individual level, alternative psychological assumptions on preferences or rationality may be introduced to explain why individuals cooperate (see Gächter 2013). Another approach is to maintain

the standard assumptions that actors are rational and selfish, and instead look for features of *social* conditions that make the emergence of cooperation possible (see Buskens and Raub 2013). This is the approach typically taken by sociologists, following Coleman's (1987) suggestion that sociological theory ought to keep assumptions on individual behavior as simple as possible, in order to be able to study complex mechanisms on the social level in more detail. Nevertheless, also theoretical work that is primarily interested in social mechanisms often requires careful consideration of micro-level assumptions. Several well-known conditions that can facilitate cooperation included individual actors being involved in repeated interactions (Axelrod 1984), actors organized in social networks (Raub and Weesie 1990), and institutional arrangements that facilitate trust or cooperation (Greif 2006). Many authors in this volume build on one of these lines of research for explaining cooperative behavior in social dilemmas.

Part I of this book collects advancements in theoretical work in this area. In this part, some authors focus more on the essence of the trust or cooperation problems, while others concentrate more on the psychological assumptions and social conditions that can foster trust and cooperation. In addition to and in synergy with this theoretical work, researchers have invited people to the social science laboratory and have had them actually play these and similar kinds of games, trying to figure out whether or not the predictions that theoretical social science has come up with, fit with the behavior of people under strict laboratory conditions. We showcase work that fits this tradition in Part II. Finally, some research ventures outside the lab and confronts the predictions of cooperation theory, especially the ones that seem to hold under controlled conditions, with the empirical reality of everyday life. We show this kind of work in Part III.

The combination of, on the one hand, careful consideration of micro-assumptions using formal theoretical reasoning, and, on the other hand, testing of hypotheses under controlled laboratory conditions and under blurry real life conditions, is both a strength of the research on cooperation problems and a source of further complexity. Using mixed-methods with different strengths and weaknesses offers a potentially more robust set of explanations for societal phenomena (see Levitt and List 2007; Buskens and Raub 2013; Jackson and Cox 2013; Raub 2017). Rigorous or even formal theory ensures that assumptions are made explicit and concrete. Experimental testing can, as much as reasonably possible, try to fix characteristics of the interaction so that the observed behavior can be tested relatively independent of interfering factors. Field studies then assess whether the ideas about which conditions govern behavior are strong enough to survive in a noisy setting. The connection between these levels, however, is not always obvious. Ideas about individual preferences or behavior do not necessarily automatically translate to obvious or desired consequences at the collective level (Coleman 1987). The assumptions underlying the formal theory can be easily contended to be too strong abstractions and unsupported tests of predictions in both experimental and field studies leave open where to improve. Should theory

be improved, perhaps using more realistic assumptions? Is the empirical test not warranted under these conditions, or did the prediction simply massively fail? It is this interplay of individual-level modeling combined with more and less rigorous testing that we seek to demonstrate in this volume.

1.2 Theoretical contributions

The theoretical contributions in this volume all fit within the general theoretical framework of methodological individualism and follow Coleman's (1987) theoretical approach that emphasizes the importance of specifying micro-macro transitions in sociological theory. Within this general theoretical framework, the theoretical elaborations can roughly be divided into three categories.

The first category could, with a nod to Coleman's (1990) seminal work, be described as being concerned with the *foundations of sociological theory* on cooperation problems. The first two contributions focus on the non-trivial and much-debated question of selecting the appropriate model for individual decision making, representing the micro level of Coleman's (1987) celebrated meta-theoretical scheme. **Voss** evaluates the use of assumptions of rational egoism (related to the Hume-Buchanan doctrine) in institutional design and argues that, while non-standard assumptions (such as assumptions on bounded rationality) may be useful in institutional *analysis*, institutional *design* is still better served by the standard assumptions of rational egoism. **Opp** discusses the extent to which micro-level models of human decision making known as dual-process theories – in particular the MODE model (Fazio and Olson 2014) and the Model of Frame Selection (Esser 1990) challenge more conventional rational choice theory, and concludes that such theories complement rather than contradict rational choice theory. While the selection of micro-level models has received much attention in the literature, **Van Assen and Dijkstra** draw, in the last chapter of this first category, attention to the macro side of Coleman's famous advice that sociological theory should simplify mostly at the micro level but as little as possible at the macro level. Using the case of exchange theory as an illustration, they argue that the issue of "sufficient complexity" is too often neglected in sociological theorizing, leading to models that are oversimplified at the macro level and therefore lack ecological validity.

The second category of theoretical contributions focuses on a specific way of trying to solve the social dilemma structure: the role of reputation in the emergence of cooperation and trust. This topic has been central to rational choice-oriented sociology ever since Granovetter's (1985) programmatic paper on the "problem of embeddedness" and Raub and Weesie's (1990) first game-theoretic elaboration of that program. **Flache** builds on the seminal work by Raub and Weesie (1990) on embeddedness effects on cooperation, extending the analysis to "large number dilemmas" (Raub 1988), in particular collective good production under uncertainty. Focusing

on core-periphery networks, the theoretical analysis suggests that in such networks situations may emerge in which the core is exploited by the periphery. **Lindenberg, Wittek and Giardini** challenge the view that human cooperation can be explained by reputation effects based on rational choice and selfishness assumptions alone ("Granovetter's error"), and argue that reputation effects cannot be studied without also considering the dynamics of normative embeddedness. **Van de Rijt and Frey** build on their earlier work on *reputation cascades*, a previously overlooked dynamic of reputation systems (for instance occurring in online markets) by causing arbitrary inequality in payoffs between trustees (Frey and Van de Rijt 2016). They show by means of computational methods that their initial findings are robust against the relaxation of two restrictive assumptions of the earlier model, namely, that information is transferred automatically and reliably and that interaction sequences are relatively short. **Flap and Ultee** reflect upon the effectiveness of "organized distrust", or the extent to which actors in powerful positions are systematically scrutinized to prevent them from exhibiting opportunistic behavior.

The third category contains theoretical contributions that deal with cooperation more broadly and model effects of information exchange and networks based on well-specified micro-assumptions. **Hegselmann** studies two currently much-debated social phenomena – polarization and radicalization of opinions – as extensions of the bounded confidence model. Using agent-based modeling, he shows that for both types of phenomena bridges are of crucial importance, but also that results may depend strongly on small differences in parameters, highlighting the complexity of micro-macro transitions in these phenomena. **Bojanowski** considers individual benefits of network positions, building on Burt's (1992) structural holes model. Using an agent-based model as well, he shows that informational properties of collaborative relations should be analyzed not only at the node or tie level, but also at the triadic level to appreciate the full spectrum of possible redundancies in information. He ends with suggestions for empirical work to test his theoretical conjectures. **Gautschi** proposes a new theoretical model in the domain of network exchange that, contrary to existing sociological models of network exchange, systematically takes into account that negotiation partners pursue their self-interest and thereby specifies the actors' optimization problem more explicitly.

The consistency among the theoretical chapters, despite the broad range of substantive topics addressed, illustrates the usefulness of a common meta-theoretical approach. In particular, it shows the ability of Coleman's macro-micro-macro framework to force researchers to be explicit about what the individual-level assumptions are and to take the often complex link between the micro and macro level seriously. Any categorization of chapters is to some extent artificial, but the underlying consistency of the theoretical chapters is also illustrated by the fact that one could easily arrive at alternative categorizations. For example, besides the more foundational chapters, most of the chapters are in one way or the other concerned with micro-level assumptions in the theories discussed, with the chapter

by Lindenberg et al. providing the most obvious bridge between the "foundations" and "reputation" categories. Similarly, the chapter by Gautschi not only relates to Van Assen and Dijkstra's chapter by criticizing existing theories of network exchange (although from a different angle), but also makes an explicit effort to draw network exchange theory closer to the theories typically applied in the literature on cooperation and reputation by incorporating elements of non-cooperative game theory. Also, the contributions by Flap and Ultee on the one hand and Voss on the other are concerned with issues of institutional design. Finally, and perhaps most obviously, in many of the theoretical contributions the notions of social networks and embeddedness play an explicit or implicit role.

1.3 Experimental tests

Part II consists of chapters that consider experimental tests related to social dilemma problems. All chapters can once again be linked to Coleman's macro-micro-macro framework. A first category of chapters focus more on the micro level, studying the role of individual preferences, identities and priming. The second category of chapters zooms in on how framing of the individual choice situation can affect behavior in social dilemmas such as the previously mentioned Prisoner's Dilemma and its multi-player variant, the Public Goods Game. The last category of chapters in this part uses experiments to test how social and institutional embeddedness might solve cooperation problems.

Two chapters fall in the first category. **Aksoy** studies an extension of social value orientation theory for the situation in which there are ingroup and outgroup others, as opposed to the more standard assumption that preferences are such that all others are considered equivalent. He finds, using an experimental test based on the Decomposed Game, that many subjects show ingroup bias in the sense that they add an extra negative weight to outcomes of the outgroup when the outgroup is better off than the ingroup. **Winter and Diekmann** replicate and extend an experiment by Vohs, Mead, and Goode (2006) in which it was argued and shown that framing individual decisions explicitly in terms of a money frame matters for decisions of participants. However, the same kind of priming does not lead to differences in behavior in several strategic situations, such as the Ultimatum Game, Trust Game, Prisoner's Dilemma, and Volunteer's Dilemma. This contrasts with results that are observed using other priming manipulations (Liberman, Samuels, and Ross 2004), highlighting the theoretical puzzle we alluded to above: under which circumstances are more psychological additions to standard game-theoretic arguments needed to understand behavior in strategic situations?

Abraham, Lorek and Prosch is the first chapter that focuses on micro-level assumptions in social dilemma games. They study how a frame of being a member of a

group affects behavior in Public Goods Games and Chicken Games. They distinguish between treatments in which the frame is without any further obligations and treatments in which the frame implies a minimum contribution or a cost to enter. They show that, in both games and under different experimental circumstances, framing is effective and increases cooperation. When the cooperative frame comes at a cost the effects tend to be smaller, but only in one experiment this difference is significant. **Esser** provides an insightful comparison of the application of different versions of rational choice theory as well as of the model of frame selection in explaining behavior in Public Goods Games with a punishment option. He argues that the micro-level assumptions related to frame selection are superior in understanding these behaviors compared to different versions of rational choice assumptions, especially given the behavioral patterns when actors change between conditions with or without punishment. This is another chapter that illustrates the challenge of including psychological aspects for understanding cooperative behavior, but at the same time shows its potential for further steps in this direction given careful reconstruction of existing experimental work.

The last three chapters in this part discuss different forms of institutions in relation to different types of cooperation problems. **Engel and Ockenfels** evaluate an interesting instrument to prevent cooperative behavior in a situation in which this is societally undesirable, namely, collusion between suppliers in a market with few suppliers. Using an experiment resembling a Cournot market, they contend that in markets in which collusion between competitors is likely, the introduction or existence of an actor with competitive social preferences (a "maverick") puts the collusion between firms under pressure. Therefore, it seems a good strategy to protect mavericks in markets in which cartels are easily formed, but undesirable from a societal view point. **Corten, Buskens, and Rosenkranz** start from a more standard rational choice perspective and test experimentally in which scenarios conditions for cooperation seem to be more beneficial, crossing conditions in which reputational information about behavior in Prisoner's Dilemmas can be exchanged with conditions in which subjects can choose with whom to play these dilemmas. They cannot confirm that the availability of networks through which reputation can spread promotes cooperation in this case. However, they do find that learning effects play a role in the sense that initial cooperation levels affect cooperation in the long run. They also find that partner choice alone, without possibilities for reputational information spread, jeopardizes cooperation. These findings suggest that standard rational choice models do not suffice to understand how network related mechanisms affect cooperation in exogenous and endogenous networks. **Barrera, Buskens, and De Rover** study the effects of positive and negative sanctions in repeated Public Goods Games. They show that punishments have a stronger effect on cooperation than rewards. They also show that group solidarity is not lower in groups in which punishments could be given compared to groups in which rewards could be given. Apparently, the level of cooperation that is reached in a group is

more important for group solidarity than whether the sanctions to reach this are positive or negative.

The results of this last chapter in Part II again illustrate links between the chapters in this book. The findings of the last chapter suggesting that standard rational choice assumptions do not suffice to understand cooperation in dynamic networks illustrate that the efforts of the earlier chapters in Part II, on which individual-level assumptions seem more empirically valid, are also important for the more complex networked games. The chapter links as well to some of the theoretical contributions in Part I studying the effects of more complex assumptions in networked social situations on micro and macro-level outcomes.

1.4 Field studies

Part III considers a variety of contexts in which the mechanisms related to trust and cooperation are empirically tested in the field, underscoring that these mechanisms are applicable in areas as diverse as organizations, households, neighborhoods and networks of adolescents (cf. Raub and Weesie 2000). The first three chapters focus on the organization of bilateral relations, between firms and within households. The last two chapters in this part inquire into networked settings such as neighborhoods and classrooms and the consequences of these networks.

Rooks, Snijders, and Tazelaar consider how organizations create the rules that govern their (contractual) interactions. They contrast the view that rules are rational adaptations to the given circumstances with the view that rules are merely irrational coincidences. Their empirical analyses suggest that existing rules tend to lead to more elaborate ex ante investments, and do not necessarily improve eventual exchange performance. Apparently, the classical Weberian view that rules and rule-following are crucial elements to be able to reap the benefits of specialization through division of work needs additional argumentation about the emergence of these rules, as organizational evolution does not seem to automatically converge to the optimal sets of rules. This once again highlights the potential of incorporating micro-assumptions that are descriptively more accurate into frameworks that are based on rational profit maximization. **Koster** continues the search for empirical evidence on how organizations solve problems of trust and cooperation by considering whether organizations that collaborate on Human Resource issues are more innovative than those who do not. It turns out that organizations that do collaborate are more innovative. This reveals that, next to more standard approaches that emphasize compatibility of resources, shared cooperative endeavors can have positive consequences on organizational innovativeness as well. **Roeters, De Ruijter, and Van der Lippe** turn our attention to households, where similar trust and cooperation arguments play a role as between firms, albeit at the interpersonal rather

than interorganizational level. Their analysis on the provision of informal care shows that, even though the content and subjective experience of human relations are completely different from those in organizational interaction, theories of trust and cooperation are usefully applicable and can shed light on the conditions under which collective goods problems are solved. Informal care is less extensive when there are more coordination problems, through for instance the combination of busy day-jobs with many other roles. An interesting puzzle remains, however, as those with more general skills spend less rather than more time in informal care, highlighting the complicated issue of how to link the micro-level behavior of humans to the outcomes at the aggregate level.

Volker shifts attention to a level in between that of personal and organizational relations: the collective efficacy of neighborhoods. Similar to organizations, neighborhood members choose, given a certain level of trust between them, the rules that define the formal and informal control on behavior in neighborhoods. This in turn determines how well the neighborhood is dealing with the provision of its collective goods, such as the availability of mutual help and a lack of unwanted behavior. In this sense, there is an obvious parallel between this chapter and that of Rooks et al. Although the relation between trust, control, and collective efficacy is theoretically demanding, the results suggest that trust and control are only moderately related across Dutch neighborhoods. Moreover, trust seems to be related to collective efficacy more than control is. Especially the fact that trust and control do not seem to reinforce each other, is an interesting avenue for future theorizing and empirical research. **Snijders and Kalter** switch back and forth between micro-foundations and macro testing in a study on cooperative ties and social cohesion in classrooms. Making use of the Stochastic Actor-Oriented Model (Snijders and Steglich 2015), they tackle the empirical puzzle that the number of mutual within-classroom friendships is declining as a function of the proportion of Muslims in a class, whenever Muslims are a minority. Successive iterations of increasingly realistic models hint at an explanation that is based on the idea that friendship ties between ethnically diverse students are more sticky, although the models cannot completely explain the phenomenon away. We see here that slight variations in micro assumptions matter, as well as that the transition from micro behavior to macro-phenomena, once again, is far from trivial.

Summarizing, Part III shows that field studies can illustrate which theoretical predictions consistently replicate in different contexts. In addition, most studies also produce new puzzles that require further theorizing using alternative assumptions about micro-level decision making and more elaborate inclusion of macro-level conditions. This requires further theoretical studies and different types of empirical tests of the new theories to establish whether the new explanations are more encompassing than the earlier ones.

1.5 Conclusion

This collection of chapters not only shows a way in which research on trust and cooperation problems can and perhaps even should progress, but also how rigorous sociology in general could progress more consistently. It is a combination of careful theorizing, when necessary at the individual level, to figure out how individual preferences can lead to behavior under different circumstances, combined with a core of rigorous model building that can help ensure that theorizing is consistent and extendable. We then see that the link from micro to macro level often requires explicit theoretical arguments to include the necessary interdependencies between individuals and reveal unintended consequences of individual behavior. In addition, the meticulous testing of arguments under first ideal, but progressively realistic circumstances to compare the predictions with actual behavior, and finally the tests of general insights at the intricate level of reality where everything is related to everything else. If predictions fail, theory needs to be updated, ideas amended, and conditions become more specific. This mix of methods can help to overcome a tendency of getting tied up into prohibitively local problems and to keep an open mind for the struggles, solutions, and victories of researchers in seemingly other, but actually related areas.

References

Axelrod, Robert 1984. *The Evolution of Cooperation*. New York: Basic Books.

Burt, Ronald S. 1992. *Structural Holes. The Social Structure of Competition*. Cambridge, MA: Harvard University Press.

Buskens, Vincent, and Werner Raub. 2013. "Rational Choice Research on Social Dilemmas." Pp. 113–150 in *Handbook of Rational Choice Social Research*, eds. Rafael Wittek, Tom A.B. Snijders, and Victor Nee. Stanford, CA: Stanford University Press.

Coleman, James S. 1987. "Microfoundations and Macrosocial Behavior." Pp. 153–173 in *The Micro-Macro Link*, eds. Jeffrey C. Alexander, Bernhard Giesen, Richard Münch and Neil J. Smelser. Berkeley, CA: University of California Press.

Coleman, James S. 1990. *Foundations of Social Theory*. Cambridge, MA: Belknap Press of Harvard University Press.

Dasgupta, Partha. 1988. "Trust as a Commodity." Pp. 49–72 in *Trust: Making and Breaking Cooperative Relations*, ed. Diego Gambetta. Oxford: Blackwell.

Esser, Hartmut. 1990. "'Habits', 'Frames' und 'Rational Choice'." *Zeitschrift für Soziologie* 19(4): 231–47.

Fazio, Russel H., and Michael A. Olson. 2014. "The MODE Model: Attitude-Behavior Processes as a Function of Motivation and Opportunity." Pp. 155–71 in *Dual-Process Theories of the Social Mind*, eds. Jeffrey W.S Sherman, Bertram Gawronski, and Yaacov Trope. New York: Guilford Press.

Frey, Vincenz, and Arnout van de Rijt. 2016. "Arbitrary Inequality in Reputation Systems." *Scientific Reports* 6:38304.

Gächter, Simon. 2013. “Rationality, Social Preferences, and Strategic Decision-making from a Behavioral Economics Perspective.” Pp. 33–71 in *Handbook of Rational Choice Social Research*, eds. Rafael Wittek, Tom A.B. Snijders, and Victor Nee. Stanford, CA: Stanford University Press.

Granovetter, Mark S. 1985. “Economic Action and Social Structure: The Problem of Embeddedness.” *American Journal of Sociology* 91:481–510.

Greif, Avner. 2006. *Institutions and the Path to the Modern Economy: Lessons from Medieval Trad*e. Cambridge: Cambridge University Press.

Hechter, Michael, and Satoshi Kanazawa.1997. “Sociological Rational Choice Theory.” *Annual Review of Sociology* 23:191–214.

Hedström, Peter, and Peter Bearman (eds). 2009. *The Oxford Handbook of Analytical Sociology*. Oxford: Oxford University Press.

Jackson, Michelle, and David R. Cox. 2013. “The Principles of Experimental Design and Their Application in Sociology.” *Annual Review of Sociology* 39:27–49.

Kollock, Peter. 1998. “Social Dilemmas: The Anatomy of Cooperation.” *Annual Review of Sociology* 24:183–214.

Levitt, Steven D. and John A. List. 2007. “What Do Laboratory Experiments Measuring Social Preferences Reveal About the Real World?” *Journal of Economic Perspectives* 21:153–174.

Liberman, Varda, Steven M. Samuels, and Lee Ross. 2004. “The Name of the Game. Predictive Power of Reputations Versus Situational Labels in Determining Prisoner’s Dilemma Game Moves.” *Personal and Social Psychology Bulletin* 30:1175–1185.

Luce, R. Duncan and Howard Raiffa. 1957. *Games and Decisions*. New York: Wiley.

Raub, Werner. 1988. “Problematic Social Situations and the ‘Large-Number Dilemma’.” *Journal of Mathematical Sociology* 13:311–357.

Raub, Werner. 2017. *Rational Models*. Utrecht: Utrecht University.

Raub, Werner, and Vincent Buskens. 2011. “Aujourd’hui, une Sociologie Rigoureuse.” *Commentaire* 136:1063–1066.

Raub, Werner, Vincent Buskens, and Rense Corten. 2015. “Social Dilemmas and Cooperation.” Pp. 597–626 in *Handbuch Modellbildung und Simulation in den Sozialwissenschaften*, eds. Norman Braun and Nicole Saam, Wiesbaden: Springer.

Raub, Werner and Thomas Voss. 2018. “Micro-Macro Models in Sociology: Antecedents of Coleman’s Diagram. Pp. 11–36 in *Social Dilemmas, Institutions, and the Evolution of Cooperation*, eds. Ben Jann, Wojtek Przepiorka. Berlin: De Gruyter.

Raub, Werner, and Jeroen Weesie. 1990. “Reputation and Efficiency in Social Interactions: An Example of Network Effects.” *American Journal of Sociology* 96:626–654.

Raub, Werner, and Jeroen Weesie. 2000. “The Management of Matches: A Research Program on Solidarity in Durable Social Relations.” *The Netherlands Journal of Social Sciences* 36:71–88.

Snijders, Tom A. B. and Christian Steglich. 2015. “Representing Micro-Macro Linkages by Actor-Based Dynamic Network Models.” *Sociological Methods & Research* 44:222–271.

Vohs, Kathleen D., Nicole L. Mead, and Miranda R. Goode. 2006. “The Psychological Consequences of Money.” *Science*, 314:1154–1156.

Part I: **Theoretical Contributions**

Thomas Voss

2 Institutional Design and Human Motivation: The Role of Homo Economicus Assumptions

Abstract: What kinds of behavioral or motivational assumptions are appropriate if "legislators" want to design "good" social institutions or constitutions? David Hume's famous advice has been to follow the maxim "that every man must be supposed a knave: Though at the same time, it appears somewhat strange, that a maxim should be true in *politics*, which is false in *fact*." Notice that Hume as well as Adam Smith both argued that humans sometimes and under certain conditions empirically act in accordance with a principle of "sympathy", that is, they are able and willing to take the roles of their interaction partners and identify with their respective interests. Nevertheless, legislators who try to construct efficient social institutions should not assume that such pro-social motives prevail. In contemporary constitutional economics James Buchanan has endorsed Hume's maxim with regard to the design of basic societal institutions: "*Homo economicus,* the rational, self-oriented maximizer of contemporary economic theory, is, we believe, the appropriate model of human behavior for use in evaluating the workings of different institutional orders" (Brennan and Buchanan 1985: 61). The paper evaluates the Hume-Buchanan doctrine in the light of empirical evidence and theoretical insights with respect to social preferences and intrinsic motivations.

2.1 Introduction: Explaining institutions versus. designing institutions

Institutions are "rules" which affect outcomes of various social interactions. The set of institutions which regulate interactions in modern societies comprises a variety of rules on different levels. There are informal and formal rules like conventions which regulate every-day interactions in traffic situations. Other rules refer to informal norms which call for contributions to public goods ("avoid leaving refusals

Note: Material from this paper has been presented at the Seminar "Rational Choice Sociology: Theory and Empirical Applications" at Venice International University, San Servolo, November 21 - November 24, 2016. Comments by participants of this Seminar and by two referees, in particular by Hartmut Esser, are gratefully acknowledged.

Thomas Voss, Department of Sociology, Leipzig University, Germany

https://doi.org/10.1515/9783110647495-002

when having a picnic on the green"). More formalized rules are laws and constitutions which have been designed consciously. Online auction platforms like eBay use rules which refer to the auction procedure (e.g., variants of a second-price auction) and to improve incentives to cooperation (e.g., reputation mechanisms) which have been designed and which are continuously monitored and improved. At the multi-national level core actors from different states (e.g., from the European Union) currently design, evaluate, interpret and change rules which are intended to solve severe collective action problems with respect to financial and monetary policy, international migration and others.

Social science and sociology in particular aim to explain the workings, the emergence and effects of institutions. There has been and still is of course a lot of controversy about the theoretical tools and methodology which is appropriate in this context. Durkheim (1895) who defined sociology as the science that has to describe and explain social institutions and other social facts claimed that social facts should be explained by empirical laws on the level of social facts. In this paper, I will adopt, in contrast, the methodological individualistic idea that explaining institutions (that is, social facts on the "collective" or "macro" level) requires micro level assumptions about individual behavior or choices. In some contexts it may be useful to model individual actions by applying ideas which resemble Popper's "logic of the situation" (Popper 1966) or in some other "everyday sense" of rationality and not in the technical sense of axiomatic rational choice theory (Simon 1978). In economics, political science and sociology, however, more technical rational actor assumptions and models were employed with considerable success. The micro model of man until the 1990's has predominantly and more specifically been a version of the homo economicus or rational egoism (RE) conception. There are many variants of this conception but as the two basic ingredients I consider the assumptions of rationality (R) and self-interestedness or "egoism" (E). These ingredients are conceptually and logically independent: R basically means consistency of preferences (and beliefs) and E means that a particular subset from the extensive set of (logically consistent) preferences is empirically relevant. In other words, social preferences (e.g., altruism, fairness, spitefulness) can be and are in fact presently used in a rational actor model but not in the homo economicus or RE concept. The RE model is the core explanatory component in Coleman's (1990) opus magnum as well as in much of Gary Becker's (1976) seminal economic approach to social behavior. The RE approach is considerably strong in demonstrating the collective effects of social structural and institutional variables. In sociology, there has been an extensive research program based on game-theoretic principles about the structural factors which foster trust and cooperation in "problematic" social dilemma situations. In much of this work, individual behavior is modeled as driven by self-interested motives and by instrumental rationality which means that behavioral choices are assumed to be consistent with (possibly subgame perfect) Nash equilibria. In social dilemmas there is a contrast between rationality and efficiency

(in the Pareto sense) among rational egoists. It has been shown by Werner Raub and members of his Utrecht research group that certain social conditions (like repeated interaction with a large "shadow of the future", multiplexity, multilateral reputation and others) may contribute to efficiency gains in social dilemmas (see Raub and Weesie 1990, Raub et al. 2013, Raub 2017; other closely related work in a similar vein is described in Greif 2006; for experimental work that demonstrates fairly good empirical predictions of RE assumptions in the repeated prisoner's dilemma see, for example, Dal Bó 2005). However, as is well known, there has been something akin a scientific revolution within economics and several related fields which casts serious doubts on the empirical and explanatory status of RE assumptions. Early attacks on the RE model focused on the rationality assumption (e.g., transitivity and independence axioms in expected utility theory). In more recent decades laboratory experiments and behavioral game theory apparently demonstrated that (many or some) subjects are motivated by social preferences and not exclusively by self-regarding preferences. For the time being, there seems to be a consensus that empirically valid explanations of institutions must be based on or at least be complemented by behavioral and motivational assumptions which are psychologically more "realistic" than the RE model in that they employ bounded rationality and social (and possibly also intrinsic) motivations (see Ostrom 2005a: Chapter 4; Bowles 2016). However, there is less of a consensus about which particular version of bounded rationality or of social preference functions should be used.

This very rough sketch of the state of the art refers to *empirical explanations*. The situation may look different if *normative* questions of *institutional design* are considered. In his well-known Presidential Address to the American Sociological Association Coleman argued in favor of sociology as a science of design and rational construction of society:

> What does this (. . .) mean for sociology and sociologists? It implies a future in the design of organizations, institutions, and social environments – design intended to optimize relevant outcomes. (. . .) It is the task of sociologist to aid in that construction, to bring to it the understanding of social processes, to ensure that this reconstruction of society is not naive, but sophisticated, to ensure, one might say, that it is indeed a rational reconstruction of society.
>
> (Coleman 1993: 14)

Many social scientists still accept Max Weber's postulate of a value-free social science ("Werturteilsfreiheit"). According to Weber's postulate social science cannot legitimately give an ultimate justification for value judgments or normative recommendations. But scientific methods can help to investigate whether two or more normative statements are mutually consistent. Social science can also, and this is a task of eminent practical importance, demonstrate whether and by which means certain normative goals can be realized. With respect to designing institutions, normative social theory may have the legitimate task to provide what has been called

"hypothetical imperatives" (Harsanyi 1958). In contrast to unconditional categorical imperatives, they contain conditional statements of the form "If you accept certain rationality axioms or if you want to realize certain goals then you should choose alternative X".

In fact, the basic idea of some prominent neo-contractarian approaches to normative social theory is to use Gedankenexperimente of a "moral decision situation" (Harsanyi) or an "original position" (Rawls) behind a "veil of ignorance" to argue which rules or which principles of justice rational individuals will choose. In addition to such "constitutional" decisions, normative theory might also (and in fact: must) analyze the expected effects of those principles or rules given they will be implemented in a future society or group. Such analyses must be based on assumptions about the social conditions which prevail in the situation as well as the behavioral principles which are at work among the (future) participants of the society. It seems likely that the theory of action which is employed in analyses of constitutional decisions and in analyses of agents' behaviors in a future society will not necessarily be identical. To be more specific, with respect to constitutional decisions one may arguably use rationality principles of the RE type, whereas with regard to post-constitutional decisions behavioral principles of the bounded rationality type can be appropriate.

In this paper, I will discuss the theoretical role of RE assumptions in the context of normative social theory. The main task will be an evaluation of what may be called Hume-Buchanan doctrine of using homo economicus or RE assumptions in institutional design. Prior analyses of the doctrine (Schüssler 1988, Hausman 1998) mainly focus on different critical aspects which will not be covered in the present paper.

2.2 The RE model and the Hume-Buchanan doctrine

2.2.1 Three levels of decisions in institutional design

In referring to normative social theory, I will confine myself to subjectivist and consequentialist approaches in the following sense. In the tradition of David Hume's moral philosophy or in accordance with Max Weber's idea of a "Verantwortungsethik" normative problems are necessarily explicated by pointing to the subjective interests of the individuals who are the target actors and the beneficiaries of normative or moral imperatives. In the light of this tradition it must be demonstrated that the choice of and the conformity to normative principles is in fact possible and will have consequences which are consistent with the long-term subjective interests of the involved agents.

It is convenient to distinguish between three levels of decision problems in normative social theory. First, (**level 1**) there are first order decisions with respect to the *criteria* institutions should fulfill: Efficiency, maximizing "social welfare", "wealth", just redistribution, profit maximization etc. Secondly, (**level 2**) there are second order decisions with respect to the choice of *specific* rules or institutions which are consistent with the criteria (*constitutional choice*). These rules are designed such that they provide incentives and constraints for the involved (future) agents to realize first order goals. Thirdly, (**level 3**) one has to deal with third order decisions under the constraints of the level 2 rules. In this case, the "social planner" has the task to analyze individual decisions and collective outcomes of the (fictitious) agents who (in the future) will live under the rules which are implemented in accordance with first and second order decisions (if so). Notice that decisions on levels 1 and level 2 are, so to speak, decisions on a "meta level" which are due to the actors who want to design an institution. Level 3 decisions are on the "object level" which are, however, dependent on the "planner's" selection among models of man.

2.2.2 Example: Institutional design in migration policy

To illustrate with a concrete case, consider the example of institutions which regulate refugee and migration policy in the European context. To regulate immigration into the European Community, Betts and Collier (2017) evaluate various institutional solutions (see also Collier 2013). They argue (level 1) in favor of the normative criterion of welfare maximization as a principle that is consistent with utilitarianism. (The set of individuals whose utility levels are to be maximized contains not only prospective migrants but also the people who remain in African or Asian countries ["stayers"] and European citizens. This is a kind of cosmopolitan normative criterion because the welfare (or the utility levels or "happiness") of people from foreign countries is included – not only Europeans' welfare). Evaluating alternative rules (level 2) requires the demonstration of how the rules will affect the dominant actors' incentives (level 3). These agents are (i) political leaders who are faced with collective action problems and incentives to free ride on the supra national level ("let other countries contribute to the costs"), (ii) refugees searching for individually optimal options and (iii) agents ("entrepreneurs") who are interested to engage in smuggling as an (illegal) commercial enterprise.

As another example, I mention reflections on institutional design of immigration from an US American perspective (Posner 2013). As the maximand on level 1 Posner argues in favor of the well-being of Americans. The welfare of American citizens (workers, employers, other citizens) depends on the individual characteristics of the migrants who are selected and who possibly enter the labor market. Social welfare may also be positively affected by migrants whose presence permits family reunification. Furthermore migrants who are willing to integrate or to assimilate

the cultural and social norms and who thereby will increase (or at least not decrease) society's social capital may be preferred to migrants who are not willing or able to integrate. Migrants of the "good" type contribute to an increase of American welfare, "bad" types do not bring valuable human capital or other good characteristics. In the extreme, bad types, once they have been admitted to stay, may look for a career as beneficiaries of public welfare or even for a criminal career. There are various institutional solutions to *screen* potential migrants in order to select those persons who are "good" types. The screening methods have to account for asymmetric information problems and reduce the adverse selection and moral hazard problems which can arise if migrants are better informed about their own type than the government agencies that are deciding about admission of migrants. In addition to screening potential migrants, there must also be, as argued by Posner, rules and procedures to *control* the behavior and rights of migrants. As will be clear to anyone who has thought about the design of institutions in this realm, the rational construction of efficient rules which serve the selected normative goals is extremely complex and cannot be discussed in detail in the present paper. It is, however, clear that analyses on level 2 and level 3 (second and third order decisions) are critically interconnected. To base screening predominantly on educational credentials and other desirable qualifications (language proficiency etc.) that can be easily observed (as is done in the Canadian points system) can have the adverse effect to deny admission to migrants who may be potentially highly motivated unskilled laborers – provided that the country would profit from admissions of some migrants who are unskilled workers. Another problem is that a points system does not help to cope with the control or moral hazard problems *after* admission has been permitted (see Posner 2013: 303–304). Posner's analysis is based on the assumption that potential migrants are heterogeneous with respect to the distribution of desirable characteristics and that their behavior is rational and motivated by self-interest under the constraints of the rules which regulate admission to the United States. The line of reasoning that is sketched in Posner's contribution is completely consistent with a standard economic (RE) model of man.

2.2.3 Designing "just" or morally "good" institutions

Given that the outcomes of institutions should meet certain (normative) criteria, social theory can define its task as the specification and evaluation of institutions. There are of course well-known attempts to address this task with respect to the basic constitutional institutions such that these institutions meet criteria of moral reasoning. In moral theory, Harsanyi and Rawls, among others, asked which kinds of general principles and rules meet moral standards of fairness or impartiality. This problem refers to first order and second order decisions. Harsanyi and Rawls both adopted a rational actor approach, albeit in slightly different ways. The actors

are envisioned in a Gedankenexperiment as agents behind a "veil of ignorance" such that they choose rules *as if* they did not knew which social position they will occupy within the (future) society that will realize concrete institutions under the constraints of the chosen universal principles. Harsanyi (1976, 1977, 1978) argued in favor of Bayesian rationality (expected utility maximization) and demonstrates that rational agents in a "*moral decision situation*" will necessarily choose institutions which maximize the aggregate welfare as measured by the sum (or arithmetic mean) of the cardinal utilities of the individuals. This approach of course requires that there are theoretically sound methods to interpersonally compare individual utility levels. The normative principle which emerges as an outcome of Harsanyi's moral decision situation is consistent with utilitarian ethics, in particular "rule utilitarianism". Harsanyi's approach is silent about the concrete institutions which help maximize utilitarian welfare functions. Rawls (1971), in contrast, argued that agents will be extremely risk averse in the "original position" of this constitutional choice situation. They will therefore, according to Rawls, use the maximin criterion of decision theory. This line of reasoning and some additional arguments build a justification for Rawls' so-called "difference principle" which implies that just institutions should guarantee that – if social inequalities cannot be prevented or would be inefficient in the Pareto sense – society allocates primary goods such that the welfare level of incumbents of the most disadvantaged positions is maximized (compared to the elements of the set of alternative institutions). Rawls' approach consequently is less abstract than Harsanyi's in that it points to more concrete institutions (like equal opportunity in education) of modern societies which work in the direction of realizing the goals which are expressed in the principles of justice. This, however, comes at the cost that it is doubtful whether the relevant agents on level 3 of third order decisions in fact will want to or will be able to enforce the required policy recommendations (for instance with respect to the organization of educational systems) (see Coleman 1974 for an elaboration of this critique).

In a similar vein, *constitutional economics* as developed in the Public Choice School asks which basic constitutional rules rational agents behind a "veil of uncertainty" would choose. The rules which are chosen refer to situations of collective action and rules of collective decision-making in some future situation. As a case in point think about individuals who decide about whether to allocate economic resources by markets or by bureaucratic organizations. Persons who presently are successful entrepreneurs may of course have different personal interests than members of a public organization that is developing or enforcing laws to regulate market behavior. The latter expect better career prospects if the amount of bureaucratic regulation is strengthened. The rationale of a veil of uncertainty is to secure a certain degree of impartiality such that the decision is devoid of immediate self-interest. In Buchanan's approach "persons are modeled as though they were faced with choices among rules of social order that are generally applicable and guaranteed to be quasi-permanent. By comparison, the Rawlsian 'veil of ignorance' is an

idealized normative construction, the appropriate starting point for persons when they consider making choices among basic principles of justice" (Brennan and Buchanan 1985: 35). Choosing among rules behind a veil of uncertainty means that "the interest of any person or group is much less easily identified" than in the choice within the set of rules (in the post-constitutional decisions). This is so because "rules are, almost by definition, applicable to a number of instances or cases" and because rules "embody an extended time dimension. The very notion of a rule implies existence through a sequence of time periods" (Brennan and Buchanan 1985: 35). As an example, Buchanan considers balanced-budget rules which restrict democratic governments' discretion with respect to public debt. Buchanan argues that the amount of conflict of interest among agents choosing *among* basic constitutional rules (level 1 and level 2) will be much lower than in decisions *within* the set of rules (level 3). Agreement will therefore be reached more easily and often unanimously.

2.2.4 Three levels of decisions in designing "just" or morally "good" institutions

To reiterate, there are three analytically distinct levels of decision problems which are involved in institutional design. On all three levels the choices of possibly multiple heterogeneous agents must be modeled (members of a constitutional assembly, directors of an enterprise, members of a political party etc.). In contractarian approaches to morality and in constitutional economics, the multiplicity of personal interests and their potential conflicts are reduced by imposing the fiction of a veil of ignorance or uncertainty. It may be argued that the veil will potentially (empirically or in light of abstract rationality) generate a consensus or a unanimous decision of different agents. The decisions on this level may then be analyzed by treating the decision-making body as a single representative individual. Given the restrictions of a veil, the decisions are considered in these approaches as consistent with rationality axioms of Bayesian decision theory and game theory (Harsanyi) or neoclassical economic theory (Buchanan) or a particular non-Bayesian conception of individual rationality (maximin) with extreme uncertainty (Rawls). In Harsanyi's work (1977) which contains a precise explication of the concept of a moral decision, each involved agent (on level 1) is uncertain about her future social status or position in society and assigns an equal probability to each outcome (in terms of status positions). She then chooses the set of rules that maximizes her expected utility (in terms of von Neumann-Morgenstern utility functions). There is no need to assume altruism or other social preferences at this point because even perfectly selfish individuals *under the constraints of the veil* will choose rules which maximize the social welfare as defined by utilitarian criteria (sum of personal utilities of all involved individual positions).

It is, however, debatable whether and under which conditions rational self-interested individuals will want to *enter a decision situation behind a veil* of ignorance or uncertainty and will want to commit themselves to accept the set of rules which have been selected before in the constitutional stage. As many critics have emphasized it may even be debatable whether or not real humans will in fact attain a consensual decision on this level (see, for example, Coleman 1974, 1990). Thus, on level 1 it is assumed that behavior (behind a veil) is governed by principles of rationality in the RE sense (or homo economicus sense). Referring to second and particularly third order decisions Buchanan's constitutional political economy is most explicit. Arguments on level 3 require that the social scientist has to compare the empirical effects of various institutional arrangements with respect to the desired outcomes. This means that appropriate assumptions with respect to actors who choose within the set of rules must be employed. It is important to note that decision-making on level 1 has to take into account second and third order decisions. Table 2.1 gives an overview on the three levels and some illustrations.

Table 2.1: Three levels of choices in institutional design.

	Decision-making agent(s)	**Constraints**	**Set of alternatives**
First order (level 1) decisions	Subject(s) who design institutions: – general citizenship – constitutional assembly – political decision-making body	"Moral situations": veil of ignorance (Harsanyi, Rawls) "Constitutional choices": veil of uncertainty (Buchanan); Collective decision-making rules; Other physical or social constraints and resources	Universal principles of justice; Maximands of individual and or social welfare
Second order (level 2) decisions	Subjects who design or implement institutions	Universal principles and criteria which institutions should fulfill	Concrete realizations of first order choices: institutions
Third order (level 3) decisions	Subjects who live under the constraints of institutions	Institutions as realized outcomes of first and second order choices	Opportunity set under the constraints of institutions

To illustrate, designing rules with respect to immigration policy not only requires decisions about the normative criteria and goals which should be reached but must also include predictions about the likely effects of these rules given the cognitive

and motivational capacities of the various involved agents who will act under the institutions. Thus all three decision levels are interconnected.

2.2.5 The Hume-Buchanan doctrine

Referring to institutional design Buchanan and Brennan approvingly quote David Hume's famous advice to follow the maxim "*that every man must be supposed a knave*: Though at the same time, it appears somewhat strange, that a maxim should be true in *politics*, which is false in *fact*" (Hume 1741: 42–43). Notice that Hume as well as Adam Smith both argued that humans sometimes and under certain conditions empirically act in accordance with a principle of "sympathy", that is, they are able and willing to take the roles of their interaction partners and identify with their respective interests. Nevertheless, legislators who try to construct efficient social institutions should not assume that such pro-social motives prevail. In contemporary constitutional economics Buchanan and Brennan endorsed Hume's maxim with regard to the design of basic societal institutions as follows: "*Homo economicus,* the rational, self-oriented maximizer of contemporary economic theory, is, we believe, the appropriate model of human behavior for use in evaluating the workings of different institutional orders" (Brennan and Buchanan 1985: 61). Interestingly, this doctrine is justified not primarily by the *empirical* "realism" or adequacy of the model of man but by using rationality behind the veil: Since we are *uncertain* with respect to properties (in particular preferences and information) of the relevant actors who in future instances will act under the restrictions of the constitutional rules we should use the decision rule of a "*quasi risk aversion*" when deciding among various models of man: "Our claim is that because of the nature of what is to be evaluated, the gains attached to an 'improvement' secured by departures of behavior from the modeled are less than the losses imposed by corresponding departures of behavior in the opposing direction, that is, toward behavior worse than that represented in the model itself" (Brennan and Buchanan 1985: 63). The rationale for this normative maxim is not that human behavior is empirically governed by rational egoism under all circumstances but that even in populations with a majority of actors who are endowed with pro-social motivations these motives may be driven out: "the narrow pursuit of self-interest by a subset will induce all persons to behave similarly, simply in order to protect themselves against members of the subset" (Brennan and Buchanan 1985: 68). In other words, if institutional rules are designed under the premise of pro-social behavior of the target actors (who are supposed to follow the rules) there is a risk to trigger a crowding out process with respect to these motivations ("Gresham's law of politics").

2.3 Anomalies and the Hume-Buchanan doctrine

2.3.1 The RE model of man and its anomalies

It may be useful to classify models of man along two dimensions: Rationality and motivation (preferences). The homo economicus or RE model is first characterized by (1) selfish motivation. This means that actors are outcome-oriented and self-regarding. Outcomes of behavior may be non-material such as units of social approval and status received in social exchanges. (2) The second dimension is (the degree of) rationality. Neoclassical price theory assumes complete and consistent preferences and perfect information-processing abilities. In Bayesian game theory many other capacities are built into the model: behavior in accordance with subjective expected utility maximization axioms (of Savage), common knowledge of rationality, backward induction.

To be more specific, rationality means that agents choose among alternative courses of action in accordance with certain basic assumptions and more specific criteria (rationality axioms). The most general background assumptions are consequentialism and invariance. *Consequentialism* (in the descriptive sense) in this context means that actions are selected in accordance with (expected) future consequences. This rules out the consideration of "sunk costs" or other features related to investments in the past as far as they are irrelevant for future outcomes. *Invariance* means that alternatives are evaluated and selected irrespective of how these alternatives are presented ("framing").

Standard rational choice theory, and homo economicus assumptions in particular, have of course been targets of severe criticism for being empirically inadequate. Behavioral economics and cognitive psychology have produced a great deal of empirical evidence, mostly laboratory experiments, which seem to demonstrate limitations of RE assumptions, at least on the micro-level of individual choice behavior. It has been questioned that this evidence is consequential for the analysis of real-world problems outside the laboratory (Levitt and List 2007, 2008). However, some evidence from field experiments and case studies suggests that the set of behavioral assumptions which are useful in empirical explanations of institutions must be more comprehensive than the standard homo economicus concept (see Ostrom 2005a, b).

According to prospect theory, preferences (as represented by a "value function") are situation dependent. Alternative amounts of wealth are evaluated as changes of wealth. If the change is "framed" as a "gain" compared to a reference point (status quo), the curvature of the value function is concave ("risk aversive"); "losses" are evaluated by means of a convex function ("risk-seeking"). In addition to framing, prospect theory also deals with other anomalies, e.g. problems related to a basic axiom of expected utility theory, viz. the independence assumption. The famous Allais paradox demonstrates that preferences over risky prospects are not linear in probabilities (Machina 1987); prospect theory postulates preferences and probability functions which cope with this problem (see Kahneman and Tversky 1986).

Behavioral economics pointed to several further problems of the RE model. Consequentialism has been shown to be violated by evidence on sunk cost effects (see Thaler 1980, 2015). There are also some other behavioral anomalies, some of which can be explained by prospect theory (e.g., endowment effect).

With regard to motivational assumptions, behavioral game theory seems to demonstrate that the assumption of self-regarding preferences is empirically questionable, at least a considerable proportion of participants in game experiments are acting as if they had *social preferences*. Numerous experiments on ultimatums, prisoner's dilemmas and linear public goods reveal that many subjects do not choose actions or strategies such that they maximize monetary payoffs (as homo economicus would).

In addition to social preferences which, for example, depend on outcomes of members of ego's reference group (like altruism, inequity aversion or envy) and not only on ego's own material payoffs, there is an extensive body of work on *"intrinsic" motivation*. Humans may be motivated to perform certain actions because these actions are valuable per se. Under certain conditions intrinsic preferences (for example, to cooperate) will be crowded out if external interventions provide material rewards or punishments. There is anecdotal and experimental evidence from psychology and also, but less so, from experimental economics which demonstrates such crowding out effects. One can argue that in cases of intrinsic preferences a central assumption of neoclassical economics is violated, namely additive separability (Bowles 2016). In the context of intrinsic preferences this means that externally provided incentives and intrinsic motives to perform an action do not work additively but there may be interferences such that external incentives on the one hand *increase* the tendency to choose the relevant action and on the other hand *reduce* the strength of intrinsic motives. The *total effect* may be in the extreme such that the propensity to perform the desired action is reduced due to the incentives – contrary to the goals which the "social planner" wants to accomplish. As a case in point consider the effects of intensive external monitoring or variable rewards (piece-rate payments) on workers who have been, prior to the intervention, intrinsically motivated to perform their tasks. In such cases, it may well be that interventions decrease the quality and quantity of workers' outputs because they undermine intrinsic motivations.

An overview on selected anomalies is given in Table 2.2.

Table 2.2: Empirical anomalies of the RE model of man.

RE Assumptions	Anomalous effects	References (sample)
I. Rationality (R)		
Consequentialism (Outcome-orientation)	Sunk cost effects	Thaler 1980; Thaler 2015: Chapter 8
Invariance	Situation dependent preferences: Framing	Kahneman & Tversky 1986; Kahneman 2011

Table 2.2 (continued)

RE Assumptions	Anomalous effects	References (sample)
Independence Axiom	Allais paradox	Machina 1987
II. Motivation (E)		
Selfish Preferences	Behavioral Game Theory: Social preferences (Ultimatum, Public Goods with Punishments experiments)	Camerer 2003
Separability of extrinsic and intrinsic rewards	Intrinsic Motivation: Crowding out- effects of material incentives	Frey 1994; Bowles and Polania-Rayes 2012; Bowles 2016

2.3.2 The RE model is not the "worst case"

2.3.2.1 Anti-social motives

With respect to *motivational assumptions* is the RE model in fact, as the Hume-Buchanan doctrine supposed, the "worst case" assumption? The idea is that different social institutions which regulate peculiar "problematic" social situations can be analytically matched with various distributions of individual preferences (self-regarding vs. social preferences of various kinds). Then various combinations of preferences with institutions generate different outcomes which can be evaluated normatively. Now, not every logically possible combination of individual preferences is empirically relevant. Some authors indeed have argued that the set of individual preferences can analytically be reduced to those utility arguments which are shared by every human being and related to a stable subset of basic commodities (see Becker 1976). However, some experiments reveal that indeed altruism, fairness but also anti-social preferences are effective. Yet, for the time being, we do know next to nothing about the empirical distribution of preferences in various situations. In some social contexts laboratory experiments have evoked an emotional reaction of spitefulness and costly punishment against target actors who contributed to public goods (Herrmann et al. 2008). If those motives prevail, the opportunity to punish free riders (and possibly also cooperators) in a public goods situation yields inefficient outcomes. It thus seems that self-interestedness is not always the worst case.

2.3.2.2 Choice anomalies

With respect to behavioral anomalies which reflect *limits of rationality* it is tempting to ask whether the complete rationality assumption is indeed a worst case. Consider

an anomaly like the *sunk cost effect*. An agent who knows that she is not immune to such effects when she deals with everyday operational decisions but who wants to maximize her personal welfare will want to commit herself to rules which prevent her from trapping into sunk cost fallacies. These institutions would serve to improve outcomes of RE's in terms of their material interests. This means the homo economicus model is not the most pessimistic ("worst case") model because anomalies can yield individually and/or collectively suboptimal outcomes. The general point thus is as follows: Homo economicus is not the kind of person whose behavior is prone to behavioral anomalies. However, ordinary people who want to realize self-regarding preferences will want to live under the constraints of rules which cope with their limits of rationality. In fact, the very idea of "liberal paternalism" and "nudging" (Thaler and Sunstein 2008) is to improve everyday decisions (the criterion of improvement is the realization of RE preferences). But this implies first and second order decisions which are by and large consistent with the RE model. Thus, the nudging approach is a conception of institutional design which rests on the idea that rules should be superimposed on boundedly rational agents who want to attain better outcomes in terms of self-interested preferences.

2.3.3 Crowding out effects: Gresham's law of politics versus other crowding out effects

A central argument that is used to support the Hume-Buchanan doctrine is based on what Brennan and Buchanan (1985: 68) call Gresham's law of politics. Suppose a set of rules produces "good" or efficient outcomes provided a large fraction of the population of actors who act under the constraints of these rules are endowed with pro-social preferences. Then it is likely that a crowding out process takes place such that pro-social motives gradually disappear from the setting provided self regarding actors are more successful in this situation. This being the case, it seems that rationality requires as a "quasi risk averse" choice of the model of man to use the RE model in analyses of third order decisions. Though there is some supporting evidence for this kind of social dynamics (see, however, Schüssler 1990 for some caveats) other types of crowding out processes have been studied extensively.

In addition to crowding out processes, it may also be the case that crowding in-processes take place. For instance, if a critical mass of pro-social agents is clustered within a population of RE agents and if the pro-social agents (much) more frequently interact with one another than with the remaining population, they may in long term drive out the self-regarding agents because they are more successful. This process has been described in various contributions to evolutionary dynamics, for example group selection approaches (see Gintis 2017: Chapter 9 for a discussion of these mechanisms based on a debatable adoption of group selection ideas).

Psychological mechanisms which lead to driving out processes are another important case in point. Given that an agent is intrinsically motivated to perform certain desirable actions, it may well happen that external interventions via the provision of extrinsic material rewards or threats of punishments will drive out the motivation to a considerable degree. External interventions perceived as *controlling* can reduce intrinsic motivation. There are also crowding *in* effects: In some cases intrinsic motivations may be fostered by interventions which are perceived as *supportive*. Effects of this kind have been studied by psychologists experimentally with children who reduced their effort to perform certain tasks after ostensive observation and or material incentives had been introduced. More recently, some evidence from experimental games seems to support these effects (for a review and explication of some experimental results see Bowles 2016: 39–77). It seems safe to say that institutional design based on the provision of material incentives has to account for such crowding out processes (for an extensive overview of the literature see Underhill 2016). Depending on the particular parameters of the situation the provision of extrinsic incentives can result in an absolute reduction of intrinsic motives within the individual. In cases of severe reductions of intrinsic motivation, the total effect of both, intrinsic and extrinsic incentives, may be such that the desirable behavior is repressed. In other words, the imposition of rules via external incentives can be counterproductive provided that a significant proportion of the population is endowed with intrinsic preferences which favor the desired behavioral outcomes.

2.3.4 Social preferences and institutional design

There is some evidence that social preferences exist even in market contexts. Online auction platforms are certainly products of conscious institutional design. The online transaction company eBay uses not only auction mechanisms which have been adopted from mechanism design theory which is based on the RE model of man. But in addition, eBay created and implemented institutions to cope with trust and cooperation problems that arise in online transactions between partners who in general cannot rely on the shadow of the future or a shadow of the past because exchange is not repeated. The use of multilateral reputation is of course not new but has been standard in commercial relations in economic history since a long time (Greif 2006). eBay uses a feedback system such that buyers can evaluate their partner after the transaction is completed and that this evaluation diffuses throughout the system. The whole community of market participants can in principle use the reputation scores attached to a prospective partner before deciding to interact with that partner. As far as we know, eBay modified its institution several times in order to make it immune against fraud. Until 2008 it was not only possible for buyers to evaluate the seller's trustworthiness but sellers could also evaluate buyers. This led to some problems due to a kind of negative reciprocity: Sellers

responded to negative feedback by retaliating and giving a negative evaluation even in cases that the buyer's behavior had been exemplary correct. This problem was mitigated via institutional reform but the feedback system still relies on motives which cannot be assumed as prevailing among RE agents. Why should an RE buyer evaluate her partner at all? Giving a feedback is profitable only for the community of future (potential) transaction partners of the focal seller. It is not valuable to the evaluator (assuming that he will not repeat a transaction with the same partner) who has to bear a low cost if thinking about and in fact giving feedback. In other words, to evaluate is comparable to contributing to a public good. Since there are in general no material incentives for giving a feedback, the workings of the reputation system depend to a certain degree on the pervasiveness of social preferences among the users of the platform. There is evidence that in fact a majority of eBay users gives feedback (Bolton et al. 2013, Diekmann et al. 2014). The general point is this: Some institutions depend on voluntary contributions to public goods and work fairly well even though there is no material compensation for cooperation. There apparently is a critical mass, if not a large majority, of participants who display behavior that is based on social preferences. Another institution that heavily depends on voluntary cooperation is of course democratic voting. Still another is participation as a respondent in survey research interviews or questionnaires.

2.4 Why RE assumptions are useful in institutional design

2.4.1 First order and second order decisions and rationality

Choice anomalies due to limited rationality and non-standard motivations are relevant and must be considered in designing institutions. This does, however, not imply that the RE model is obsolete. Consider first choice anomalies like sunk cost effects. As has been argued, these anomalies create a demand for rules that serve to help overcome the resulting individual or collective inefficiencies. The "nudging" program's aim is indeed to supply those rules: "A nudge is some small feature in the environment that attracts our attention and influences behavior. Nudges are effective for humans, but not for Econs [RE agents, T.V.], since Econs are already doing the right thing. Nudges are supposedly irrelevant factors that influence our choices in ways that make us better off"(Thaler 2015: 326). Evaluating the suboptimality of anomalous behavior is thus, according to the nudging approach, based on the RE model. Nudging therefore rests on the assumption that third order decisions are boundedly rational. The goals of institutional design (first and second order decisions) are, on the other hand, determined by applications of the homo economicus model.

2.4.2 Unclear effects of intrinsic motivation

According to the RE model agents systematically react on external incentives. Rules which change relative prices of alternatives are powerful tools for behavioral change ("power of incentives"). This is, in essence, the basic idea of the law of demand of standard neoclassical microeconomics. Assuming institutional design is based on this idea and that there are (also or entirely) agents who are (without being provided with external incentives) motivated to perform the desired action per se, there will be a chance that crowding out takes place. The effects of interventions via rules or regulation must, however, be analyzed with care: The total effect can be decomposed into the direct effects of extrinsic and intrinsic incentives plus an indirect effect (possibly crowding out). It is obvious that the direction (sign) and strength of the total effect depends on the values of each of the three parameters. It may be that an indirect effect of extrinsic incentives on intrinsic motives is strong enough to offset the direct effect which results in a negative total effect. This is the most severe crowding out effect. Other cases are logically possible, for instance, crowding *in* effects such that incentives reinforce the effects of intrinsic motivation. This may occur if incentives are perceived as supportive. The central questions are: Is there any reliable ex-ante information about whether social and intrinsic preferences will exist and persist at all? If so, what can be known about the exact distribution and strength of these motivations for different social settings? Since laboratory and field experiments have, for the time being, generated results which are somehow inconclusive or context-dependent, these questions have no clear answer.

It has been argued that evidence from laboratory studies is far more supportive for the impact of non-standard motivational factors than field experiments (Levitt and List 2008). In the latter context, the power of material incentives seems more effective. A particularly well researched area has been employment relationships and the impact of pay-for-performance or piece-rate payments. In certain contexts where one may conjecture that a significant proportion of laborers is intrinsically motivated to supply high levels of effort (e.g., teaching or research and development) pay for performance may indeed be counterproductive in that it elicits poor "quality" and yields an increase in cheating behavior (Jacob and Levitt 2003). Besides crowding out, these counterproductive effects have been mostly attributed to difficulties in measuring work quality and in providing inappropriate indicators for performance.

On the other hand, there exists supportive empirical evidence on the power of incentives in this particular context. From a famous natural experiment clear evidence emerged on increases in effort (without loss of work quality) which could be attributed to the introduction of piece-rate payment (Lazear 2000). The study population was employees of an auto glass repair corporation who prior to the intervention received hourly wages for delivering an acceptable level of output. Quality of work seemed not to be relevant because insufficient quality would be detected

immediately. There was no indication of intrinsic motivation on the side of the workers. Increases in average output per worker resulted from two mechanisms: First, individual effort was enhanced. Secondly, there was a sorting of high-effort workers who preferred to stay with the firm or were hired, whereas other workers looked for jobs at different firms.

2.4.3 Institutional domains: Market competition and corporate actors

There are thus situations where the assumption that the agents are to a significant degree endowed with social or intrinsic preferences which are related to the desired outcome (productivity, cooperation, etc.) does not seem likely. To argue more systematically, there are certain institutional domains where the relevant actors' motivation functions obviously do not depend on other-regarding or intrinsic preferences. The mechanisms which are at work in these domains are market competition, learning and the formalization of standard operating procedures which constrain agents' behavior. Consider anonymous markets without trust problems. In certain areas of the financial industry, and in financial markets in general, the RE assumption is likely a plausible and empirically fruitful first approximation to manage problems of institutional design. In recent years many commentators from the financial press complained about certain adverse phenomena and social dilemmas: With respect to computerized high speed trading of stocks and derivatives there is obviously a severe arms race analogous to a Prisoner's dilemma. Trading firms compete with each other about the fastest connection to Wall Street or other stock exchanges and invest substantial resources to become some fraction of a second faster than their competitors. This gives rise to extra profits by "sniping", that is, by exploiting informational advantages. For example, if a trader gets the information that Apple stock prices increase in New York, her automated ordering system can place orders in Chicago within milliseconds where the prices will similarly rise with a small time lag. However, this competitive advantage vanishes if other traders use similarly quick connections and automated trading software. Another problem is criminal behavior called "spoofing" which is not easily detected and punished by legal sanctions (Ford 2016). This means to place (so to speak fake) buying or selling orders which change the price of some stock. Some milliseconds before the order is scheduled to become executed the spoofer cancels the order and can collect an extra profit by simultaneously selling or buying mirror stock that he previously has acquired.

Possible interventions are the imposition of a small fee on all cancelled trades above a certain limit or to require a minimum rest time before orders to remain open. These new rules would probably not completely reduce opportunism or wasteful investments into rat races but mitigate the problems to some degree. Regulations of this kind would change the behavior of traders by the power of relative prices or

opportunity costs. Other instruments which rest on attempts to change traders' preferences, moralizing, or appeals to some code of honorable trading behavior (or the attempt to change the "culture" of the financial or investment banking sector) will as far as we know not work. There is also probably no risk that some "intrinsic motivation" to act honorable will be crowded out by such interventions.

In social theory, James Coleman (1982, 1990) has advanced the concept of "corporate actors" and of an "asymmetric society". Modern societies are increasingly dominated by corporate actors like capitalist organizations. There is an asymmetry in power in relations between natural persons and these corporate actors. Interestingly, Coleman explains corporate action by means of the RE model. However, natural persons' inferior power position is in part due to the fact that they are, according to Coleman, subject to choice anomalies, for example weakness of will and time inconsistency. These anomalies are systematically exploited by corporate actors. To illustrate, in certain business relationships corporate actors supply their customers with contracts which contain default options with respect to quitting the relation, payment or protection of data privacy. As is well known from behavioral research, opting out from a default option is chosen much less frequently than simply to accept the default option (a prominent example is rules of organ donation). This may be explained by choice behavior on the side of natural persons that is largely consistent with the RE model (costs of switching to an alternative) or, which is somewhat more intuitive, by weakness of will or other types of choice anomalies. One might say that many corporate actors have tacitly or intentionally employed some of the rules and recommendations of the nudging approach in order to further profit interests. In other words, corporate actors have rationally designed institutional rules and contracts with the goal to maximize corporate profits (first and second order decisions) under the assumption that natural persons' behavior is subject to certain choice anomalies (third order decisions).

In the spirit of Coleman's approach, it may be a legitimate goal of normative social theory to provide the public with recommendations on how to improve the power of natural persons in their relations with corporate actors. It would be beside the point to create rules on the assumption that the behavior of capitalist corporate actors (or Weberian bureaucratic organizations more generally) is governed to a significant degree by social or intrinsic preferences. There is indeed some empirical evidence that natural persons are inclined to act more opportunistically vis-à-vis corporations. Individuals tend to frame their interactions (e.g., contracts) with corporate actors much less in terms of morally neutral exchanges than interactions with other natural persons (Rai and Diermeier 2015).

2.4.4 An excursus on low cost effects

One might argue that the power of relative price changes is confined to the domain of markets and capitalist corporate actors. However, there is an intensive discussion

in rational choice sociology about the so-called low cost hypothesis in everyday environmental behavior. The debate set off from papers by Diekmann and Preisendörfer (2003). Though the focus has predominantly been on environmental behavior (by natural persons), the low cost phenomenon occurs in other contexts too where moral or other normative attitudes interact with material interests.

Starting point have been empirical findings based on survey research data:

1. Normative standards have a positive effect on pro-environmental behavior
2. Costs of pro-environmental behavior negatively affect behavior
3. Interaction effects: The positive effect of norms declines with increasing costs

Crediting Douglass North and other economists, Diekmann and Preisendörfer proposed the so-called low cost hypothesis to explain the observation of a discrepancy between normative attitudes (as measured by interview responses) and factual behavior: Words are cheap, but deeds can be very costly. Besides papers by Diekmann and Preisendörfer, many other contributors proposed theoretical explanations of the low cost hypothesis (for an overview and discussion see Best and Kroneberg 2012; Tutic et al. 2017). Some of them have used rather complicated variants of rational choice or of framing reasoning (Best and Kroneberg 2012). Tutic et al. (2017) present a parsimonious explanation based on a simple application of classical demand theory which is a slightly modified variant of a RE-model. The idea is to use a Cobb-Douglas utility function (a tool which was, by the way, also used in Coleman's [1990] analyses of the linear system of action).The utility arguments comprise two normal goods, namely a composite good measuring the amount (quantity) of pro-environmental or other activities related to some normative standard, $x_a \geq 0$, and another good $x_{\neg a} \geq 0$ which is unrelated to the attitude or normative standard:

$$u(x_a, x_{\neg a}) = x_a^{\alpha} \cdot x_{\neg a}^{1-\alpha}$$

Exponents $\alpha \in (0, 1)$ measure the strength of the attitude. It is easily shown that by maximizing this function under the conventional linear budget equation that is used in neoclassical demand theory one can derive the three empirical regularities. One of them is of course the classical *law of demand*: The larger the relative prices of behavior that is related to the attitude or normative standard the lower the demand for this behavior. Notice that the low cost hypothesis is not identical to the law of demand but also contains a hypothesis postulating an interaction effect. This interaction is also easily derived from the Cobb-Douglas-function approach. Empirical evidence indicates that the low cost hypothesis is not universally valid (see Best and Kroneberg 2012 for references). The Cobb-Douglas approach of course depends on preferences which are related to some normative standard. Given that parameter α is very low or even equals zero, there will not be any effects of normative standards at all. It is an empirical question whether or not there are conditions such that extremely high costs do not suppress "moral" behavior. In cases like

these alternative explanatory approaches based on "framing" may be appropriate (see Esser and Kroneberg 2015).

Some further insights are implied by this kind of comparative statics. Interventions which reduce the relative price of desired moral behavior will increase the amount of that behavior and vice versa. (This may be considered as a kind of crowding in effect.) One may of course use the whole technical apparatus of the Slutsky equation to describe other effects in more detail. It is furthermore possible to make the model more "realistic", for example by introducing a full income constraint in terms of costs of time (which makes sense because many pro-environmental activities will yield direct costs in terms of time and opportunity costs in terms of reduced wages or leisure time).

It is in general not necessary to introduce some more or less complicated framing model to explain the stylized low cost effects. In other words, institutional design may be based on a slightly modified RE model that accounts for specific normative or pro-social attitudes without eliciting crowding out effects.

2.4.5 The power of incentives

With respect to institutional design, the homo economicus model suggests interventions which *change the relative prices* of alternatives such that in the aggregate a socially desirable outcome will be realized. Economics appeals to the so-called generalized *law of demand* in this case: If the relative price of a good increases, the demand for this good will decrease. In other words: Individual demand functions are negatively inclined. The implementation of these interventions is, in principle, simple because restrictions can be much better manipulated than preferences. From the analysis of the low cost hypothesis, which has been sketched, it follows that relative price effects can work even in cases where the affected individuals hold certain pro-social or moral attitudes. The analysis was comprised to a single (representative) individual agent. It is, however, easily demonstrated that in the aggregate the same qualitative effects can be predicted: Aggregate demand functions which are simply derived by a summation of individual demand curves will show the same qualitative property of a negative slope. The negative slope of the aggregate demand function does not depend on specific preferences.

An important result in this context has been proved in a seminal paper by Becker (1962, reprinted in 1976). If there are n agents who are restricted to choose a bundle of goods which is on their budget line, the negative slope of the aggregate demand function even follows for "irrational" behavior of the participants. Irrationality means that choices are selected at random (impulsively) or inertial (routinely). In other words, on the macro level of aggregate behavior the power of incentives is still preserved if agents are irrational. Even when dealing with "irrational" agents, the RE model can be an appropriate instrument (for the analysis of third order decisions) in designing institutions.

Developing this point further, notice that in contrast to RE models of third order decisions the macro consequences of limited rationality models or of social preferences utility functions are by no means clear. Many critics have noted that virtually any empirical observation on the individual level can be "explained" ex post by an ad hoc behavioral model. On the other hand, for many if not most "psychological" theories of individual behavior the aggregate consequences are far from clear. It is in many instances not possible to generate inferences on the level of collective effects via applying behavioral theories. Take for example the Kahneman-Tversky prospect theory. Though this theory is explicated with relatively great precision, it is not yet possible to systematically generate insights for strategic interaction situations. The reason is that prospect theory is not easily linked to game theoretic equilibrium concepts and therefore does not contain a heuristic for micro-macro transitions. What can be done is to translate specific theoretical components of prospect theory, for example assumptions about framing on the curvature of the value function (gain frames induce concave utility functions), into a game-theoretic context (see Raub and Snijders 1997 for such an approach). Alternatives to RE models have not yet evolved into a coherent, unified and parsimonious theoretical instrument to accomplish micro-macro-aggregations. Even proponents of behavioral theory have to concede that "for comparative static predictions of aggregate behavior, self-interest models may make empirically correct predictions because models with more complex motivational assumptions predict the same outcome" (Falk and Fischbacher 2005: 183).

2.5 Conclusion: The role of RE assumptions in institutional design

When dealing with problems of institutional design, it is important to distinguish three interconnected levels of decisions. First and second order decisions comprise the selection of (normative) criteria and goals of design and the specific rules that are appropriate to meet or realize these criteria. Third order decisions refer to the "object level" of behavioral choices selected by the (future) target actors and beneficiaries of the rules. It has been argued that with regard to first and second order decisions (sometimes called "constitutional decisions"), assumptions of rationality and self-regarding preferences, possibly behind a "veil" in certain moral contexts, can be appropriate even though with respect to third order decisions other models of man may be justifiably used. Behavioral economics and cognitive psychology have produced empirical evidence and also some theoretical insight that limits of rationality and social or intrinsic preferences may affect institutional design in important ways. Individually or socially suboptimal effects which result from boundedly rational behavior (for example, sunk cost effects) create a demand for institutions

which reduce these inefficiences. This demand arises if the outcomes are evaluated (first order) by RE assumptions.

With respect to non-standard preferences, it is necessary to generate some ex-ante information on the specific distribution of preferences or other characteristics of the (future) target actors whose third order decisions are extrapolated. The acquisition of this knowledge is more likely in cases of small and stable social situations with a fixed population of interacting individuals whose characteristics are well-known and homogenous. In populations with heterogenous individuals from different cultural contexts and with considerable migration between social contexts, "legislators" will in general not be able to gather much ex-ante-information with regard to the specific mixture of preferences. This information, however, is necessary to estimate whether or not pro-social motives will be undermined to a significant degree by incentives which are provided by the constitutional rules. With regard to normative constitutional theory, the Hume-Buchanan approach towards the construction of universal rules which are valid for an extended time horizon and possibly for a population with a high degree of diversity seems still warranted. Since such constitutional choices will necessarily be made under conditions of high uncertainty with regard to the properties of the affected agents and the specific distribution of social preferences it may be a "quasi-risk aversive" choice (Brennan and Buchanan 1985) to use the standard homo economicus assumptions as a first approximation.

Some additional points which favor the use of RE assumption in analyses of third order decisions are as follows:

1. There is a large set of bounded rationality- and non-selfish motive-models with partially contradictory predictions. It seems impossible to select *one* element from this set as a theoretical tool suitable for every problem of institutional design. In other words, given that ex-ante knowledge about the properties of the target population is lacking, the use of standard homo economicus assumptions is an obvious alternative.
2. In large-scale aggregate behavior it seems, in general, that homo economicus models and many more complicated alternative theories yield very similar, if not the same, predictions. In this case: Why not use standard RE model?
3. There are institutional domains which require design principles consistent with homo economicus assumptions:
   - Competitive, anonymous markets (highspeed trading in financial markets), auctions
   - Profit-oriented corporate actors' behavior (in relations with natural persons)
4. Non-standard models are appropriate in special situations involving decisions at the margin, e.g. certain "low cost" situations. However, non-standard models offer no clear predictions about structural variables which affect outcomes of institutional design (e.g. repeated interactions, network effects) and must therefore be combined with standard models.

Institutional design is, in general, a complex task with considerable uncertainty about its possible effects. Many, if not all, attempts of conscious design are prone to generate non-intended consequences. One should keep in mind that a trial-and-error process of "piecemeal-engineering"(Popper 1966) – albeit guided by theoretical principles – will in general be needed. Starting with RE assumptions, this process of piecemeal-engineering will possibly adopt different behavioral principles referring to third order problems depending on the concrete boundary conditions of the situation. A note of caution has been expressed by celebrated institutional design expert Elinor Ostrom with these words:

> The policy of assigning all authority to a central agency to design rules is based on a false conception that there are only a few rules that need to be considered and that only experts know these options and can design optimal policies. Our empirical research strongly challenges this assumption. There are thousands of individual rules that can be used to manage resources. No one, including a scientifically trained professional staff, can do a complete analysis of any particular situation. (Elinor Ostrom 2005b: 269)

Finally, let me conclude with two *false* propositions:

1. The RE model represents the uniquely optimal model for institutional analysis.
2. The RE model is useless in institutional analysis.

References

Becker, Gary S. 1962. "Irrational behavior and economic theory", *Journal of Political Economy* 70: 1–13 (reprinted in Becker 1976).

Becker, Gary S. 1976. *The Economic Approach to Human Behavior*. Chicago: University of Chicago Press.

Best, Henning, and Clemens Kroneberg. 2012. "Die Low-Cost-Hypothese". *Kölner Zeitschrift für Soziologie und Sozialpsychologie* 64:535–561.

Betts, Alexander, and Paul Collier. 2017. *Refuge. Transforming a Broken Refugee System*. London: Penguin Books.

Bolton, Gary, Ben Greiner, and Axel Ockenfels. 2013. "Engineering Trust: Reciprocity in the Production of Reputation Information", *Management Science* 59: 265–285.

Bowles, Samuel. 2016. *The Moral Economy – Why Good Incentives are no Substitute for Good Citizens*, New Haven: Yale University Press.

Bowles, Samuel, and Sandra Polania-Rayes. 2012. "Economic Incentives and Social Preferences: Substitutes or Complements?", *Journal of Economic Literature* 50: 368–425.

Brennan, Geoffrey, and James M. Buchanan. 1985. *The Reason of Rules – Constitutional Political Economy*, Cambridge: Cambridge University Press (reprinted as Volume 10 of The Collected Works of James M. Buchanan, Indianapolis: Liberty Fund, 2000).

Camerer, Colin F. 2003. *Behavioral Game Theory: Experiments in Strategic Interaction*, NewYork: Russell Sage.

Coleman, James S. 1974. "Inequality, Sociology, and Moral Philosophy", *American Journal of Sociology* 80 (3): 739–764.

Coleman, James S. 1982. *The Asymmetric Society*. Syracuse: Syracuse University Press.
Coleman, James S. 1990. *Foundations of Social Theory*. Cambridge, Mass.: Harvard University Press.
Coleman, James S. 1993. "The Rational Reconstruction of Society: 1992 Presidential Address", *American Sociological Review* 58: 1–15.
Collier, Paul. 2013. *Exodus. Immigration and Multiculturalism in the 21st Century*, London: Penguin Books.
Dal Bó, Pedro. 2005. "Cooperation under the shadow of the future: Experimental evidence from infinitely repeated games", *American Economic Review* 95(5): 1591–1604.
Diekmann, Andreas, and Peter Preisendörfer. 2003. "Green and Greenback. The Behavioral Effects of Environmental Attitudes in Low-Cost and High-Cost Situations". *Rationality and Society* 15:441–472.
Diekmann, Andreas, Ben Jann, Wojtek Przepiorka, and Stefan Wehrli. 2014. "Reputation Formation and the Evolution of Cooperation in Anonymous Online Markets", *American Sociological Review* 79(1): 65–85.
Durkheim, Emile. 1895. *Les règles de la méthode sociologique*. Paris: Presses Universitaires de France (18th ed., 1973).
Esser, Hartmut, and Clemens Kroneberg. 2015. "An Integrative Theory of Action: The Model of Frame Selection", pp. 63–85 in: Lawler, Edward J., Shane R. Thye, and Jeongkoo Yoon (eds.), *Order on the Edge of Chaos: Social Psychology and the Problem of Social Order*. New York: Cambridge University Press.
Falk, Armin, and Urs Fischbacher. 2005. "Modeling Strong Reciprocity", pp. 193–214 in: *Moral Sentiments and Material Interests*. Edited by Herbert Gintis, Samuel Bowles, Robert Boyd, and Ernst Fehr. Cambridge, Mass.: MIT Press.
Ford, Jonathan. 2016. "'Spoofing' case highlights perils of automated trading", *Financial Times* November 13, 2016 (online version: https://www.ft.com/content/a60b4c4c-a988-11e6-809d-c9f98a0cf216).
Frey, Bruno S. 1994. "How Intrinsic Motivation is Crowded Out and In", *Rationality and Society* 6: 334–352.
Gintis, Herbert. 2017. *Individuality and Entanglement. The Moral and Material Bases of Social Life*, Princeton: Princeton University Press.
Greif, Avner. 2006. *Institutions and the Path to the Modern Economy*, Cambridge: Cambridge University Press.
Harsanyi, John C. 1958. "Ethics in Terms of Hypothetical Imperatives", *Mind* 67: 305–316 (reprinted in Harsanyi 1976).
Harsanyi, John C. 1976. *Essays on Ethics, Social Behavior, and Scientific Explanation*. Dordrecht: Reidel.
Harsanyi, John C. 1977. *Rational Behavior and Bargaining Equilibrium in Games and Social Situations*. Cambridge: Cambridge University Press.
Harsanyi, John C. 1978. "Bayesian Decision Theory and Utilitarian Ethics", *American Economic Review* 68(2): 223–228.
Hausman, Daniel. 1998. "Rationality and Knavery", pp. 67–79 in: Werner Leinfellner and Eckehart Köhler, eds. *Game Theory, Experience, Rationality; Foundations of Social Sciences; Economics and Ethics: In Honor of John C. Harsanyi*. Dordrecht: Kluwer.
Herrmann, Benedikt, Christian Thöni, and Simon Gächter. 2008. "Antisocial Punishment Across Societies", *Science* 319: 362–367.
Hume, David. 1741. "Of the Independency of Parliament", pp. 40–47 in *Essays Moral, Political, and Literary*. edited by Eugene F. Miller. Indianapolis: Liberty Press Classics, 1985.
Jacob, Brian A., and Steven Levitt. 2003. "Rotten Apples: An Investigation of the Prevalence and Predictions of Teacher Cheating", *Quarterly Journal of Economics* 118: 843–877.

Kahneman, Daniel, and Amos Tversky. 1986. “Rational Choice and the Framing of Decisions”, S. 67–94 in: Robin M. Hogarth & Melvin W. Reder (eds.), *Rational Choice*, Chicago: University of Chicago Press.

Kahneman, Daniel. 2011. *Thinking, Fast and Slow*. London: Allen Lane.

Lazear, Edward P. 2000. “Performance Pay and Productivity”, *American Economic Review* 90 (5): 1346–1361.

Levitt, Steven D., and John List. 2007. “What Do Laboratory Experiments Measuring Social Preferences Reveal About the Real World?”, *Journal of Economic Perspectives* 21(2): 153–174.

Levitt, Steven D., and John List. 2008. “*Homo Economicus* Evolves”, *Science* 319: 909–910.

Machina, Mark. 1987. “Decisions under Uncertainty: Problems Solved and Unresolved”, *Journal of Economic Perspectives* 1:121–154.

Ostrom, Elinor. 2005a. *Understanding Institutional Diversity*. Princeton: Princeton University Press.

Ostrom, Elinor. 2005b. “Policies that crowd out reciprocity and collective action”, Pp. 253–276 in: *Moral Sentiments and Material Interests*. Edited by Herbert Gintis, Samuel Bowles, Robert Boyd, and Ernst Fehr. Cambridge, Mass. MIT Press.

Popper, Karl R. 1966. *The Open Society and Its Enemies*, (Two Volumes). Fifth ed., London: Routledge & Kegan Paul.

Posner, Eric. 2013. “The Institutional Structure of Immigration Law”, *The University of Chicago Law Review* 80: 289–313.

Rai, Tage S., and Daniel Diermeier 2015. “Corporations as Cyborgs: Organizations elicit anger but not sympathy when they can think but not feel”, *Organizational Behavior and Human Decision Processes* 126: 18–26.

Raub, Werner. 2017. *Rational Models*, Utrecht: Universiteit Utrecht (expanded version of a farewell lecture as Dean).

Raub, Werner, and Jeroen Weesie. 1990. “Reputation and Efficiency in Social Interactions”, *American Journal of Sociology* 96: 626–654.

Raub, Werner, and Chris Snijders. 1997. “Gains, Losses, and Cooperation in Social Dilemmas and Collective Action: Effects of Risk Preferences”, *Journal of Mathematical Sociology* 22: 263–302.

Raub, Werner, Vincent Buskens, and Vincenz Frey. 2013. “The rationality of social structure: Cooperation in social dilemmas through investments in and returns on social capital”, *Social Networks* 35(4): 720–732.

Rawls, John. 1971. *A Theory of Justice*. Cambridge, Mass.: Harvard University Press.

Underhill, Kristen. 2016. “When Extrinsic Incentives Displace Intrinsic Motivation: Designing Legal Carrots and Sticks to Confront the Challenge of Motivational Crowding-Out”, *Yale Journal on Regulation* 33: 213–279.

Schüssler, Rudolf. 1988. “Der Homo Oeconomicus als skeptische Fiktion”, *Kölner Zeitschrift für Soziologie und Sozialpsychologie* 40: 447–463.

Schüssler, Rudolf. 1990. “Threshold Effects and the Decline of Cooperation”, *Journal of Conflict Resolution* 40: 476–494.

Simon, Herbert A. 1978. “Rationality as Process and as Product of Thought”, *American Economic Review* 68 (No. 2, Papers and Proceedings): 1–16.

Thaler, Richard. 1980. “A Positive Theory of Consumer Choice”, *Journal of Economic Behavior and Organization* 1: 39–60.

Thaler, Richard. 2015. *Misbehaving: The Making of Behavioral Economics*, New York: Norton.

Thaler, Richard, and Cass R. Sunstein. 2008. *Nudge: Improving Decisions about Health, Wealth, and Happiness*, New Haven: Yale University Press.

Tutic, Andreas, Thomas Voss, and Ulf Liebe. 2017. “Low-Cost-Hypothese und Rationalität. Eine neue theoretische Herleitung und einige Implikationen”. *Kölner Zeitschrift für Soziologie und Sozialpsychologie* 61: 651–672.

Karl-Dieter Opp

3 Rational Choice Theory, the Model of Frame Selection and Other Dual-Process Theories. A Critical Comparison

Abstract: Dual-process theories explain behavior as well as cognitive processes. They thus compete with other theories which explain, at least in part, the same phenomena. The question then is how the theories differ and which theory is to be preferred. This article focuses on the comparison of two dual-process theories with a wide version of rational choice theory. The dual-process theories are the MODE model and the model of frame selection. The wide version of rational choice theory assumes, among other things, that all kinds of motives must be considered when a behavior is explained, that beliefs matter and that individuals do what they think is best for them (subjective utility maximization). One major result of the analyses is that basic assumptions of dual-process theories in general and the two dual-process theories discussed, namely the MODE model and the model of frame selection, do not contradict RCT but complement it.

3.1 Introduction

The rise of dual-process theories in the last few decades is a challenge for other theories in the social sciences. One of those other theories is rational choice theory (RCT). This article focuses on a particular version of RCT that includes all kinds of preferences and beliefs (which may be wrong) and assumes that actors do what they consider to be best for them in the specific situation. This is the assumption of subjective utility maximization. This wide version of RCT (for details see below) is increasingly applied in the social sciences. Therefore, a comparison with other theories such as dual-process theories seems useful.

These theories come in different versions as well. The MODE model has been confirmed very well and is thus worth to be compared with RCT. In contrast to the MODE model, advocates of the model of frame selection (MFS) claim that it contradicts and is superior to RCT. It is thus useful to select this model for comparison as well.

Providing a comparison of theories is only the first step of the following analyses. It is not useful to know only differences between theories. It is further important to know how differences between theories are to be assessed. This is the major

Karl-Dieter Opp, Universität Leipzig (Emeritus), University of Washington, Seattle (Affiliate Professor)

https://doi.org/10.1515/9783110647495-003

goal of the present paper: the question is which of the theories is superior or, to formulate it more modestly, which of the theories seems more plausible. Therefore, the title of the paper is a "critical" comparison.

This article is organized as follows. We will first expose RCT and its wide version. Proponents of the MFS raise various objections that are then discussed. Next the MODE model is exposed, and its relationship to RCT is analyzed. The exposition of the MFS, its critique and comparison with RCT follows. We end with general conclusions and suggestions for further research.

Those readers who are familiar with the different versions of RCT might skip the respective sections. RCT has been outlined because there are still numerous misunderstandings about its assumptions. Furthermore, for many readers the objections discussed are so obviously wrong that they might be skipped as well. The comparative analyses of RCT and dual-process theory begin with the section "Dual-Process Theories."

3.2 Rational choice theory

There are some hypotheses that are shared by all versions of RCT. These basic assumptions are called the "general version" of RCT which is first outlined. Then the wide version is briefly described and a special version of it, value expectancy theory.

3.2.1 The general version of rational choice theory

This version makes three assumptions.[1] (1) *Human behavior is influenced by preferences*. These are goals or objectives, not attitudes. This is particularly clear in the theory of consumer behavior in economics (see any textbook in economics such Salvatore 2003). Indifference curves depict which combinations of two goods provide equal "satisfaction" (Salvatore 2003: 62). If people are "satisfied" with some good then this means that certain goals are achieved with these goods. The first hypothesis thus asserts that a behavior is chosen that is conducive to goal realization.

(2) *Constraints or behavioral opportunities influence behavior*. These are factors that more or less limit the realization of the actor's goals. These two assumptions

1 For expositions of RCT see textbooks of economics, game theory and public choice theory such as Frey 1992; Kirchgässner 2008; Gilboa 2010 or Sandler 2001. What follows is based on Opp 1999, 2019b. These articles provide a detailed discussion of the different versions of RCT and their critique. There is a large body of literature that applies RCT to explain *crime*, beginning with Gary Becker's seminal article (1968). See in particular Cornish and Clarke 2017 whose version of RCT is very similar to the wide version. See also Opp 2020.

are shared by almost all social science schools. Actually they mean that human behavior is goal oriented and influenced by the social and physical environment.

The two assumptions do not yet specify how people act when there are several behavioral alternatives. This question is answered by assumption (3): *Actors maximize their utility*. That is to say, actors choose the action that is best for them and, thus, provides the highest satisfaction. This is compatible with consumer theory in economics: the highest indifference curve is chosen that can be achieved, given the constraints (see again Salvatore 2003: 74–79). Utility thus refers to satisfaction.

3.2.2 The narrow and the wide version of rational choice theory

The general version does not impose any restrictions on the kind of goals that are admitted in an explanation of behavior. There is a *narrow version* in which only material goals and egoism (only one's own well-being is of interest) are included. In a *wide version* all possible goals are admitted. People may strive for non-material satisfactions: they may like arts or poetry; they may be interested in the welfare of others (that is to say, goals may be altruistic); they may pursue normative goals (that is to say, heed internalized norms).

This implies that RCT does not assume only *instrumental behavior*. This is often defined as behavior that is only influenced by non-normative goals (such as getting a higher income). Since normative goals are included as possible motivations, one may say that also *non-instrumental behavior* is explained by RCT (for details see Opp 2013).

In the *narrow version* of RCT the constraints or behavioral opportunities are those that really exist. However, there might be misperception. For example, committing a crime depends, among other things, upon the *perceived* likelihood of punishment which may be at odds with the actual probability. The *wide version* of RCT includes those perceived constraints. Furthermore, human cognitive limitations ("bounded rationality") are taken into account as behavioral constraints.

Another assumption of a *narrow version* is that actors behave so that they reach the best possible outcomes, from the viewpoint of an omniscient observer. The *wide version* assumes that individuals do what they think is best for them in the specific situation. Thus, subjective and not objective utility maximization is assumed.

It is sometimes held that the wide version is *circular*: motives (especially non-material goals) and beliefs are allegedly "inferred" from the behavior. This is a gross misunderstanding. In every explanation, *the empirically relevant conditions must be determined empirically*. For example, in a survey the respondents might be presented with interview questions about their internalized norms or their beliefs (such as their expectations of punishment when they commit a crime). Statistical analyses can then determine to what extent the theory is confirmed.

This does not imply that an application of RCT to explain some phenomena such as a changing crime rate always begins from scratch. There are usually many

previous applications in the same field that have found incentives a researcher may build upon. For example, in explaining voting it has been found that a norm to vote and normative expectations of important others influence voting. The assumption is that conformity to the expectation of important others is a positive incentive. In criminology, perceived punishment or informal positive or negative rewards from others matter for committing a crime. Whether those incentives are relevant in new research, must nonetheless be empirically determined. Thus, RCT – as any other theory in the social sciences – is "practically empty" in the sense that initial conditions are not part of the theories. This is meaningful because the initial conditions are different in different situations and can thus not be included in a theory which consists of general statements. Nonetheless, researchers can use existing research as heuristic guidelines for finding incentives.

The theory, be it the narrow or wide version, *does not assume that individuals perform only deliberate behavior* (see already Becker 1976:7). The assumption is that human behavior is governed by preferences and constraints. Individuals may act spontaneously or they may consciously weigh advantages and disadvantages of behavioral alternatives. As will be seen below, RCT can be applied to explain when which form of behavior is chosen.

3.2.3 Value expectancy theory as a variant of the wide version of rational choice theory

A variant of the wide version is *value expectancy theory* (VET), also called SEU theory ("SEU" for "subjective expected utility") or simply utility theory. This theory is originally formulated by social psychologists (and not by economists where VET is called SEU or EU theory) and often applied by advocates of RCT to explain sociological phenomena.[2] The wide version of RCT just hypothesizes that preferences and beliefs influence behavior. VET specifies in greater detail *how* preferences and constraints influence behavior and how individuals decide.

VET assumes that the behavior that is performed is one of the *perceived behavioral alternatives*. Which alternative is chosen depends on the *utilities* (U) and *subjective probabilities* (p) of the *perceived behavioral consequences*. For each consequence, the utilities and probabilities are multiplied. This means, that the effect of one variable depends of the value of the other. If, for example, the subjective probability that a consequence occurs is zero, the utility does not have an effect.

2 See, for example, Feather 1982, 1991. For the history of SEU theory see Stigler 1950a, 1950b. For political science and an application to voting see Riker and Ordeshook 1968; 1973: 45–77. It is important that in the following a social-psychological version of SEU theory is described that is akin to the wide version of RCT.

To illustrate, assume a person considers at a certain time to vote or abstain from voting. Let the person expect rewards from friends when he or she participates. Another reward may be that the preferred party wins: the individual considers the extent to which his or her vote influences (p) the winning of the preferred party. Finally, assume that due to an internalized norm voting is regarded as an obligation. Two behavioral equations summarize these hypotheses. Each equation models the SEU (subjective expected utility) of a behavioral alternative:

1. SEU(vote) = p • U(rewards from friends) + p • U(preferred party will win) + + p • U(conformity to the norm to vote)
2. SEU(not vote) = p • U(rewards from friends) + p • U(preferred party will win) + + p • U(conformity to the norm to vote)

Both equations list the same behavioral consequences. It is assumed that the utilities U of the consequences are equal, regardless of the behavior chosen. But the probabilities p may differ. For example, the likelihood of getting rewards from friends may be high if the individual votes, it may be zero in case of abstaining.

So far the SEU of an action is only *defined*. Nothing is explained so far. The *theory* posits: *If the SEU of a behavioral alternative i is higher than the SEU of any other alternative j, the behavioral alternative i is chosen.*

VET is compatible with the wide version of RCT. The utilities are the preferences, the subjective probabilities refer to the (perceived) constraints or opportunities. The hypothesis that the behavior with the highest SEU is chosen means that individuals do what they think is best for them in the specific situation, that is to say, they maximize their subjective utility.

Our example indicates that the wide version is clearly *falsifiable*. The p's and U's can be measured, and multivariate analyses can show whether the predictions of the model are correct. In consumer theory (see any economics textbook such Salvatore 2003: 57–86) the highest possible indifference curve is preferred that can be realized with the given constraints. One can measure the satisfaction with goods and the perceived constraints.

It is important that SEU theory does not assume deliberate behavior. Instead, a behavior may be performed spontaneously. For example, a person will always buy a certain product without deliberating.

3.2.4 A terminological note

The term *incentives* refers to the set of preferences and constraints (or utilities and subjective probabilities of the behavioral consequences). In the voting example, the incentives for voting are the terms of the right-hand side of equation 1 and 2.

Benefits are the positive, *costs* the negative incentives. Sometimes "costs" also refer to constraints only.

3.2.5 Some untenable objections – views from advocates of the model of frame selection

Advocates of the MFS aim at showing the superiority of their theory by pointing out weaknesses of RCT. It is often not clear what version of RCT their target is when they criticize "the" theory of rational action. Their target seems to be the narrow version because their major criticisms do not hold for a wide version. If the wide version is mentioned, its assumptions are sometimes not described correctly. This procedure comes close to a *straw man strategy*: a questionable version of a theory – a narrow or misrepresented wide version – is criticized, instead of a wide version that is largely accepted by the scientific community.[3] In what follows it is argued that the major objections against "the" theory of rational action are not valid for a wide version. Since some of these objections are held not only by proponents of the MFS, the following discussion is of general interest. Furthermore, it is claimed that the MFS is not contradictory to the wide version, but complements it.

(1) The transition from the narrow to the wide version is a "degenerative problem shift" (Esser 2017: 506). The original version of RCT, as it is used by Adam Smith and the Scottish moral philosophers, does not include any restriction on the *kinds* of preferences and constraints that may explain behavior. Only later such restrictions were introduced (especially in neo-classical economics). Removing such restrictions and using RCT again as a general behavioral theory (as the wide version does) increases the validity of the theory and allows a wider application (for example, to situation where non-material incentives matter). If empirical research indicates that in general a wide range of preferences and that perceived constraints (beliefs) matter, then this is logically consistent with the general theory. There is therefore no "degenerative" problem shift. Dropping narrow assumptions (that the original version does not include) to increase the validity of the theory is rather a "progressive" problem shift: the theory is improved.

3 For example, in discussing research findings by Tversky and Kahneman (1981) Esser (2017) attacks rational choice theory without mentioning that a wide version does not make the assumptions he criticizes. See further Esser's (2018) account of "the" theory of rational action which raises the question who the advocates of such a version are. See my detailed discussion of this paper in Opp (2019c). In particular, it is shown that Esser's devastating critique of "the" theory of rational action does definitely not apply to the wide version. Esser's major argument that RCT cannot explain certain experimental findings by Fehr and Gächter (2000) is clearly untenable for the wide version. Esser's application of the model of frame selection illustrates the major flaws of this model.

It is ironic that the MFS includes the wide range of incentives admitted in the wide version as well and applies SEU (see below). The MFS might thus also be regarded as a "degenerative problem shift."

(2) RCT is false because it does not explain the origin or activation of preferences and constraints and neglects the "definition of the situation." Every theory consists of independent variables that are regarded as given. Explaining these independent variables is an extension of the theory. Such an extended theory has again independent variables that are regarded as given. Claiming that independent variables should *always* be explained would amount to an *infinite regress*. Nonetheless, one might regard it as an important task to explain certain independent variables, that is to say, to address the causes of the causes.

This implies that *hypotheses about misperception or about activation of beliefs or goals do not falsify RCT*. Those hypotheses *explain* which incentives have an impact on the behavior. For example, if a situation leads to activating wrong beliefs, then these wrong beliefs determine behavior. Thus, whatever the *definition of the situation* is, this does not contradict RCT. The situation only influences the values of the incentives.

(3) Any factor can be included in a rational choice explanation. Hedström and Ylikoski (2014: 6) hold: "Finding a rational choice model that fits a particular phenomenon becomes almost trivially easy as there are no real constraints on preferences and beliefs that can be attributed to the individuals in question." According to Kroneberg and Kalter (2012: 82), the wide version of RCT "is able to assimilate almost any psychological concept or theory and translate it into more or less 'soft' incentives or a more or less inaccurate belief." Kroneberg (2014: 111) adds that the wide version is "therefore of little explanatory power and heuristic value." It is striking that none of the authors provides a detailed argument that justifies their critique. The following analysis shows that this critique is clearly mistaken. Let subject S have two preferences or goals:

G1: The goal of flying to the moon,
G2: The goal of not getting wet.

Let S further have two beliefs:

B1: Apples are healthy,
B2: An umbrella protects against rain.

Now assume that we want to explain why *S performed U*, that is to say, brought his umbrella to his office at a certain day (this example is taken from Hedström 2005: 99–100). Is it compatible with RCT, that G1 and B2 can arbitrarily be selected to explain U? If S is patient of a mental hospital, this could be a valid explanation: S might perceive that bringing the umbrella (B2) would lead to the realization of the

goal to fly to the moon (G1). But our "normal" actor S will perform U *in order to* reach his goal G2 (not to get wet), and the belief that putting up an umbrella will protect against rain (B2) is perceived *to be instrumental for* goal realization. The general assumption that underlies this argument and RCT is that *human behavior is goal oriented*. In other words, behavior is enacted that is perceived by the actor to reach certain goals.

It is thus definitely not arbitrary which goals and beliefs are to be selected to explain a behavior. RCT has clear rules specifying how action, preferences and beliefs are related. These are *relational hypotheses* that are usually not formulated explicitly because they are so obvious.

These relational hypotheses can be formulated in the following way: (1) The action is chosen that leads, in the perception of the actor, to the realization of the actor's goals. (2) The action is chosen for which the actor believes that it realizes his or her goals most likely. There is thus no danger that the wide version is "immunized against empirical criticism by adding ever more utility components" (Diekmann and Voss 2004: 20, translation by KDO). The major reason why immunization is not possible is that it is to be determined empirically which incentives exist in the situation when the behavior is carried out.

These are the assumptions that are also implicitly applied by Hedström (2005: 99–100) in his example. It is particularly ironic that proponents of the MFS also use a wide version of RCT when they apply SEU theory (see below).

(4) RCT assumes only additive effects of costs and benefits (for example, Kroneberg, Yaish and Stocké 2010: 23). An interaction effect of incentives means that the effect of some incentives depends on the value of other incentives. Such interaction effects are clearly consistent with RCT. One would expect them if certain value combinations of incentives (that is to say, of multiplicative terms of independent variables) are particularly beneficial or costly. For example, assume there are strong general ("structural") deprivations (for example, there is a high inflation rate). Now let some "incidental" grievances (Hechter, Pfaff and Underwood 2016) occur such as a brutal police action. The effect of the general grievances on protest may depend on the degree of the structural grievances. People may in general already be so frustrated that a small additional grievance elicits strong protest. That is to say, the effect of one grievance (a kind of goal that is not realized) depends on the extent to which another (incidental) grievance occurs.

Another example is that a very strong norm (for example, not to steal) prompts the actor to disregard the non-normative goal (being in need of money). Thus, the extent to which a norm is followed depends on the intensity of non-normative goals and vice versa (for details see Opp 2017, also 2010). In regard to the explanation of protest with a wide version of RCT various interaction effects of incentives are theoretically derived and empirically confirmed (Kittel and Opp 2019).

RCT is thus clearly compatible with multiplicative effects of incentives. Whether such multiplicative effects occur has to be determined empirically.

(5) Explanatory content of the wide version is not higher than of the narrow version. I have argued that the wide version has a higher explanatory value than the narrow version (Opp 1999: 182). This is denied by Esser (2017:516–518). This argument is based on another formalization of the two versions (which does not convince me). The major argument in favor the higher explanatory content of the wide version is that it implies conditions for the validity of the narrow version. For example, let a person be in need of money and find a wallet with a high amount of money. Assume there is no likelihood of being detected if the wallet is not returned. A narrow version that only addresses egoistic motives will predict that the person will keep the wallet. This will not hold, according to the wide version, if the person has internalized a strong norm not to steal and has pity with the owner (that is to say, has a strong altruistic motivation). The wide version thus explains why the narrow version makes a wrong prediction. The example indicates that a theory T1 (wide version) shows the conditions for the validity of a theory T2 (narrow version). T1 can therefore not have a lower explanatory content than T2.

But even if this is denied one would nevertheless choose the wide version because it has a higher validity (which Esser admits). There is thus a trade-off: does one prefer a valid theory with a relatively low explanatory content, or an invalid theory with a high explanatory content? Usually one would prefer the valid theory.

(6) Other objections. There are other objections against RCT by advocates of the MFS that are definitely not tenable for a wide version. An especially extreme example is a recent attack by Esser (2018) against "the" theory of rational action (see the summary in Table 1: 16) which does definitely not hold for a wide version. To illustrate, it is asserted that in RCT empirical reference ("Bezug") is only formal about axioms and measurement is not required; preferences and expectations are "not observable," and short-term changes are only possible in regard to expectations (13). Numerous empirical applications of the wide version of RCT (in particular of VET) clearly contradict these allegations: utilities and subjective probabilities are (and, obviously, should be) measured and, thus, are observable, and they may change in the short as well as in the long run. Again, such changes must be measured. They are the initial conditions of the theory.

Another example for an untenable critique is that symbolic cues are "cheap talk," as Esser argues (Esser 2018: 13).[4] This means in game theory that communications

4 This article is based on a gross misrepresentation of RCT, at least of a wide version. I have written a detailed critique of this article and submitted it to the journal where Esser's article has been published (Zeitschrift für Soziologie). I will be glad to send the manuscript (which is in German) to interested readers (Opp 2019c).

cannot change payoffs. In a wide version of RCT, communications are stimuli of the environment and it is by no means excluded that they change incentives.

Such a characterization of "the" theory of rational choice raises the question why not its best available version is used to compare it with the MFS. Taking the best version is a stronger test than using a heavily flawed theory. The consequence of such a comparison is obviously that the MFS fares better. For example, it is easy to generate a contradiction by simply claiming that RCT assumes stable preferences. Of course, this makes RCT contradictory to almost every sociological theory because changing preferences are always admitted. The remainder of this essay compares, among other things, the MFS with a wide RCT and argues that the MFS is by no means superior and is burdened with serious shortcomings.

3.3 Dual-process theories

In this section we will first present some basic theoretical ideas of dual-process theories (DPTs). Next their compatibility with RCT will be discussed.

3.3.1 Some basic ideas

The name of this group of theories refers to the distinction between two "qualitatively different mental systems" which are labeled in different ways (Keren and Schul 2009: 533–534). There is an affective system and a deliberate system or, equivalently, system 1 and system 2. Accordingly, there are two mental processes: behavior may be spontaneous (or, equivalently, automatic) or deliberate.[5]

These processes have "typically" (Gawronski and Creighton 2013: 283) four characteristics in common (Table 3.1). Note that the processes and their features are all dichotomous. There are other descriptions of the two processes (see, for example, Evans and Frankish 2009, Table 1.1; Evans 2008: Table 2). The dichotomies are, for example, fast/slow, parallel/sequential, associative/rule based.

The two-system dichotomy and the related hypotheses have been criticized by several authors. Keren and Schul (2009) assert that they "lack conceptual clarity" and "rely on insufficient (and often inadequate) empirical evidence" (534). It is further an open question whether the mind consists of "one, two, or perhaps multiple systems" (534).

5 There is a vast literature on DPTs. A historical overview is provided by Frankish and Evans 2009. For general overviews see Evans 2008; Evans and Frankish 2009; Gawronski & Creighton 2013; Kahneman 2011:19–108. See further Chaiken and Trope 1999 and the successor volume by Sherman, Gawronski and Trope 2014; Evans and Frankish 2009. For a theoretical integration and the suggestion of a new model, see Mayerl (2009), summarized in Mayerl (2010).

Table 3.1: Distinguishing characteristics of spontaneous and deliberate cognitive systems.

System 1 (automatic processing)	System 2 (deliberate processing)
Unintentional	Intentional
Low cognitive resources	High cognitive resources
Cannot be stopped voluntarily	Can be stopped voluntarily
Unconscious	Conscious

Source: Based on Gawronski and Creighton (2013: 283)

In the present contribution our focus is on the explanation of behavior. Dual-process theories claim that there are two kinds of behavior: spontaneous and deliberate. Most behaviors are, however, *more or less* spontaneous or deliberate.[6] A behavioral sequence such as going to a supermarket consists partly of spontaneous behavior (such as the walking) and partly of deliberate behavior (such as choosing among several foods). The question addressed in the present article is how these different behaviors can be explained.

The literature on dual-process theories suggests various ideas that should be included in such explanations. One basic idea is that before a behavior is performed there are *situational cues* that activate attitudes, goals or beliefs that are stored in memory. These cues are of different kinds. They may be a letter of invitation for a conference or a traffic light. The cues elicit (or activate) cognitive elements. Which elements are activated depends, for example, on the accessibility of cognitive elements (see below). An important task is to specify which cues trigger which cognitive elements and which behavior.

These ideas can be applied to the voting example that was used to illustrate VET. Cues are newspaper reports about election dates, and information saved in memory when elections take place. The respective dates then activate various cognitive elements about voting.

3.3.2 On the compatibility of dual-process theories with rational choice theory

To what extent do the previous hypotheses contradict the wide version of RCT? In explaining more or less deliberate behavior RCT implies that the choice depends on the differential incentives of the behavior. This is a basic hypothesis of DPTs:

6 See in particular the review by Bargh et al. 2012, further Bargh and Ferguson 2000; Bodenhausen and Todd 2010; Deutsch and Strack 2010. Hassin, Uleman, and Bargh 2005; Wilson 2002.

"Deliberative processing is characterized by considerable cognitive work. It involves the scrutiny of available information and an analysis of positive and negative features, of costs and benefits" (Fazio 1990:89–90). Deliberation is thus costlier than spontaneous behavior.

DPTs further assume that individuals want to avoid costly situations. This implies that "considerable cognitive work" in the "deliberate mode" is unpleasant. Actors are characterized by "laziness, a reluctance to invest more effort than is strictly necessary" (Kahneman 2011: 31). A "'law of least effort' applies to cognitive as well a physical exertion. The law asserts that if there are several ways of achieving the same goal, people will eventually gravitate to the least demanding course of action" (Kahneman 2011: 35). This means that individuals subjectively maximize their utility which is exactly the assumption of the wide RCT.

This applies not only to deliberate behavior but to habits, routines or spontaneous behavior as well (for reviews of theory and research see Betsch, Haberstroh and Höhle 2002; Betsch and Haberstroh 2005). Often routines are adopted if a behavior is first chosen, based on calculation. If the actor has performed the behavior several times in certain situations and realizes that this is always the best he or she could do until important changes occur, a decision is made not to calculate anymore. Adopting a routine is thus a cost-saving device. This mechanism holds for many, perhaps most everyday behaviors.

The "law of least effort" suggests that not only overt behavior such as voting is influenced by costs and benefits, but various activities (in a wide sense) involved in cognitive processes. For example, if the goal of an actor is to cross a street as quickly as possible and wants to obey the law, he or she will *focus attention* on the red traffic light and not on other objects. If someone wants to rent an apartment that best satisfies his or her needs one will *think* about which features different apartments have, will *compare* these features, *weigh* the advantages and disadvantages and *decide* (that is to say, *form an intention*) to rent one of the apartments. The words printed in italics refer to "internal" actions. It is plausible that they are chosen because this is in the best interest of the actors.

Many other social psychological theories suggest as well that psychic processes are governed by costs and benefits. Dissonance theory implies, for example, that certain configurations of cognitive elements are unpleasant or dissonant, that is to say, costly, and that individuals prefer consonance. The "heuristics and biases" research program (Tversky and Kahneman 1974) assumes that at least some processes such as thinking are related to the economic model of man.

However, there are dissenting voices. For example, Boudon holds that accepting beliefs is not governed by costs and benefits. A detailed analysis of Boudon's arguments suggests that his position seems unacceptable (Opp 2014, 2019a).

The previous propositions of DPTs are relatively unspecific: they point to certain factors and do not specify in detail, under what conditions which effects are to be expected. There are two models that are more informative. One is the MODE

model, the other the MFS. The former has been developed by social psychologists and is well confirmed. The MFS has been developed by sociologists. As its advocates claim, it is confirmed as well (see below). The proponents of the MFS claim that it is an overarching new action theory the social sciences have tried to develop for a long time. It thus seems useful to discuss this model in greater detail.

3.4 The MODE model

3.4.1 An outline of the model

The MODE model (MM) has been developed by Fazio and collaborators.[7] "MODE" means "*M*otivation and *O*pportunity as *DE*terminants" (Fazio 1995: 257). The model specifies a process that consists of conditions that finally lead to more or less deliberate or spontaneous behavior. Figure 3.1 summarizes the model.

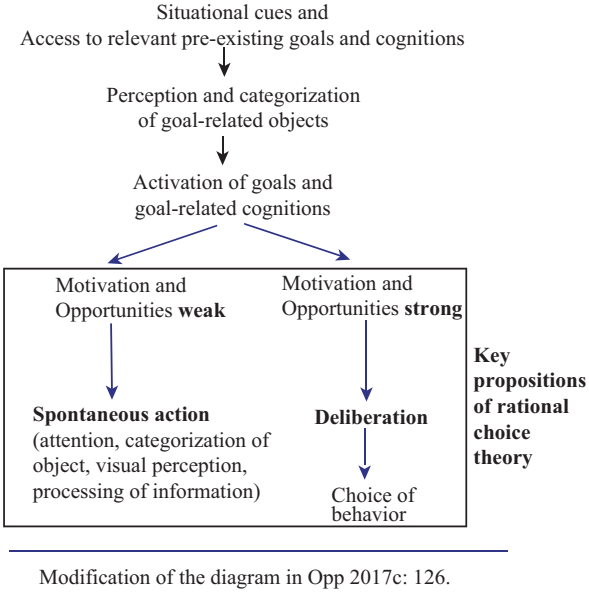


Figure 3.1: The MODE Model as a causal diagram.
Source: Modification of the diagram in Opp (2017c: 126)

7 See in particular Fazio 1986, 1990, 2001. For summaries see Fazio and Towles-Schwen 1999 and Fazio and Olson 2014. Basic measurements and tests of the model are described in Fazio 1995 and in Fazio and Olson 2014. See also the discussion in Mayerl 2009: 46–52, 78–102. In my opinion, the clearest exposition of the model is in Olson and Fazio (2009:19–27).

One of the independent variables of the model are attitudes. Social psychological research has shown that attitudes have only weak effects on behavior. There is wide agreement that goals are the major determinants of behavior (see for a detailed discussion Kruglanski et al. 2015; see already Srull and Wyer 1986). We therefore hypothesize that not attitudes but goals have effects on behavior.

In one of the first expositions of the model one might think that a sociological theory is proposed. The model begins by positing "that behavior is largely a function of the individual's definition of the situation" (Fazio 1986: 207). Fazio refers to symbolic interactionism and to William I. Thomas. The starting point of the behavioral process is that a subject gets into a situation and is thus faced with at least one attitude object in a setting. The individual then "defines" the situation. That could mean that the situation is categorized as, for example, a restaurant or a crosswalk.

Before a behavior toward an attitude object (such as a target person) can be performed, the behaviorally relevant goals and cognitions "must be accessed from memory upon observation of the attitude object" (212). *Accessibility* means that cognitions are available in memory and come "easily to mind" (Fazio 2005: 49). Only "relevant" cognitions need to be accessible, namely those that influence the behavior to be explained. If the behavior toward a target person (for example, being hostile or friendly) is to be explained, perceived attributes of the respective person are relevant and not, for example, cognitions about the politics of the US president.

The MM specifies *conditions* for the accessibility (for example, Fazio 1986: 213–215). For example, the more intense attitudes or goals are, the higher is their accessibility. For very strong attitudes, goals or other cognitions there is *chronic accessibility* (213). For limitations of space we will not deal with these effects further.

If the relevant cognitive elements are accessible they determine behavior if they are *activated* (Fazio 1990: 81). For example, if a person perceives that a traffic light is red – a certain situational cue – the object is first *categorized* as a traffic light. Then various cognitions are activated such as a norm to follow traffic rules or beliefs about consequences of breaking the rule.

It is important that only those situational cues are activated that are *related* to goals and other cognitions. In perceiving a red traffic light, the norm to stop and the belief about negative sanctioning in case of breaking the rule are activated. If a person ignores traffic lights and has the goal to cross a street if there is little traffic, the person will focus attention on the traffic and categorize the amount of traffic as more or less relevant for crossing the street. Such *relational propositions* avoid to identify the relevant cognitive elements ad hoc.

There is *selective perception*, depending on the pre-existing goals and cognitions. These perceptions influence the behavior (Bargh et al 1992: 89). If, for example, a positive attitude toward an object is activated, "then I am likely to notice, attend to, and process primarily the positive qualities of the object" (Fazio 1986: 212). Thus, selective perception is "consistent with the attitude" (213). This holds for goals as well: if the goal is, for example, to steal a wallet one will notice the

relevant qualities of a wallet (more or less filled with money) and the behavior of the owner.

Selective perception has two meanings. First, perception will be directed only at a limited number of objects in a situation. One may speak of *incomplete perception*. If a person is waiting for a traffic light to turn green, she or he will focus attention on the light and not on the kind of people waiting at the crosswalk. Second, selective perception may mean *biased perception,* that is to say, misperception. This may be the effect of intense goal states (for detailed hypotheses see, for example, Balcetis and Dunning 2006; Fazio 1986: 211; Fazio and Olson 2014: 155–157; Houston and Fazio 1989). For example, a person who is in need of money and is told that the return from an investment in Argentina is high will be more likely to believe this (that is to say, what he or she favors) than a person who is not in need of money. Ferguson and Porter (2010: 13) suggest, based on the experimental findings of Balcetis and Dunning, "that our (conscious and intentional) wants can unintentionally influence our lower level perceptions (for example, visual, auditory)." This is consistent with the everyday proposition of wishful thinking.

How the person reacts in a situation first depends on the *motivation* of the individual. This is defined as the extent to which the decision is important to an actor, that is to say, the "fear of invalidity" (Kruglanski and Freund 1983: 450). This is the costs of a "wrong" decision. For example, the decision to choose a certain college may have costly consequences, in contrast to buying a yoghurt. A low motivation is an incentive to engage in the "effortless luxury" to act without deliberation because, as is assumed, deliberating itself is costly. Actors want to avoid these costs (Fazio 1990: 89–90). Thus, motivations are in general any desires or "concerns" of an actor (Fazio and Olson 2014: 156).

A second condition for deliberation to occur is the extent to which an *opportunity* to deliberate exists. This refers to the "resources and the time for the motivated processing" (Fazio and Olson 2014: 156). "Resources" are, among other things, the cognitive capacity to deliberate and time pressure. If there is "little opportunity to engage in motivated deliberation, (. . .) judgment or behavior is likely to be influenced by the automatically activated attitude, regardless of any relevant motivational concerns" (Fazio and Olson 2014: 156). This is an interaction effect of motivation and opportunity on behavior (see below).

Motivation and opportunity are *quantitative variables*. Depending on their values one would expect *more or less deliberation and spontaneous action*. These are called mixed processes in the MM. For example, a behavior that is mainly deliberative "may still involve some components that are influenced by automatically activated attitudes" (Fazio and Towles-Schwen 1999; Fazio and Olson 2014). To illustrate, a student may scrutinize in detail the features of different colleges and then make a choice. This would be mainly deliberative. Alternatively, the student may have heard from a friend that college X is good, then checks a few features only and decides. Here less deliberation occurs. A spontaneous decision may be taken, if

the student's father and a good friend recommended college X and the student follows suit.

Both factors – motivation and opportunity – have an *interaction effect* (Fazio and Olson 2014). If there is no motivation, opportunities are not utilized. Although a retiree might have a lot of time he or she will not deliberate which yoghurt to buy.

The summary of the model in Figure 3.1 is simplified because motivation and opportunity are dichotomized. Actually, as the previous account shows, the MM implies that there may be different degrees of motivation and opportunities and, therefore, more or less deliberation.

What can the model explain? The variables the arrow points to are explananda. At the final stage of the diagram is *behavior* which may be more or less spontaneous or deliberate. The MM does not explain the origin or change of beliefs and preferences.

The voting example illustrates the MODE model. Voting may be spontaneous (a person always votes for the same party, without thinking about it) or deliberate (before the person casts his or her vote, he or she compares the party programs and then decides). If the "fear of invalidity" (the costs of a "wrong" decision) are high, there will be deliberating. This is the case if a voter thinks his or her voice matters for the outcome of an election. The opportunities are in general given: there is plenty of time for a decision, and basic mental skills for comparing parties are normally given as well. Therefore, the motivation (in the sense of the MODE model) is the decisive factor.

3.4.2 The similarity of the MODE model and rational choice theory

The MM contains the variables of RCT. *Preferences* of actors are determinants of behavior. "Motivation" refers to goals as well: it is the extent to which a decision or action satisfies the actor's goals. In particular, actors "want" to avoid useless effort of deliberating (see before).

Opportunities in the sense of time and resources limit or facilitate goal attainment. These are, as in the wide version, perceived opportunities or constraints. As the previous quotations indicate (Fazio 1990; Kahneman 2011), actors want to get the highest possible benefits, that is to say, *subjective utility maximization* is assumed (see the "law of least effort", mentioned before).

There are, however, differences between the MM and RCT. The box in the lower part of Figure 3.1 includes the part of the MM that is equivalent to RCT, as it is usually formulated. The upper part of the figure is an *extension* of (not a contradiction to) RCT. But the general assumption that also cognitive processes are based on subjective utility maximization implies that at all stages of the model actors do what they think is best for them. For example, an actor who perceives

that a traffic light is red will not activate cognitions about who will win the next Wimbledon championships.

3.5 The model of frame selection

The MFS was proposed by Hartmut Esser (for example, 1990) and further developed by Clemens Kroneberg (for example, 2006, 2014; Esser and Kroneberg 2015). This section is based upon an article by Kroneberg (2014) because this is the most recent and most extensive presentation in English. Page numbers in the text refer to this article.

The MFS has several features in common with the MM. Both models assume that cognitive elements are stored in memory. These elements are, in terms of the MFS, *mental models* or schemas (see below). In both models situational cues and accessibility are conditions for activation. Instead of "activation" the MFS uses the term "selection." For example, frame "selection" means frame "activation" (for example, 99).

Mental models of the situation are called *frames* in the MFS, whereas mental models of "sequences of actions" ("behavioral predispositions or programs of action") are called *scripts* (99). They "can refer to moral norms, conventions, routines, and emotional or cultural reaction schemes held by the actor" (99).

Now assume that an actor is exposed to a *situation*. Such a "situational object" first activates (or, equivalently, selects) a frame (which is part of the mental model). This frame then leads to the activation of a script (see, for example, Figure 4.1 and 101, see also the three formulas on p. 101). Next an action is selected. There is thus a *causal sequence* (99).[8]

Each of the selections (or activations) – frame, script and action selection – occurs in one of the following "modes": a *reflecting-calculating* (rc) and an *automatic-spontaneous* (as) mode. Activation may be more or less quick or automatic. If activation "falls below a certain threshold", that is to say, if it is relatively slow, actors "switch to the rc-mode and start to reflect about the choice in question" (101). At this point the measure of an *activation weight* (AW) for frames, scripts and actions is introduced (101). The frame, script or action with the highest AW is selected. We will only discuss the AW for a frame F_i (101) and not for scripts and actions:

(1) $AW(F_i) = m_i = o_i \cdot l_i \cdot a_i$

8 There is only one reference to *intentions*: frames and scripts "precede the building of a behavioral intention or action selection" (99). Because "intention" is never mentioned again, I will omit this variable.

The AW refers to the "immediately experienced *match* to the objective situation" (101).[9] The higher the match, the more likely is the spontaneous mode. In the example of the traffic light, the situational cue is clear, the person might have been in the situation numerous times, and the situation immediately activates the relevant cognitive elements, the behavioral program and then the action.

The three variables on the right-hand side of the equation then explain how well a frame "fits to a situation." This fit or match (empirically) depends on

> its chronic accessibility (a_i), the presence of situational objects that are significant for the frame (o_i), and the associative link between the frame and the situational objects (l_i).

In a similar way, the AW of a *script* selection is specified: the AW is high if the AW of the underlying frame and the accessibility of the script is high. The AW for an *action selection* depends on the AW of previous selections and on "the degree to which the script S_j implies a certain action" (102). There is thus a "spreading activation," that is to say, a *hierarchical process of activation.* High AWs trigger spontaneous activation and, thus, the as-mode for frames, scripts and actions. If this "process of spreading activation becomes too weak . . . the actor starts to deliberate over the perceived alternatives and makes a reflected choice" (102). The rc-mode is thus chosen.

Assume now the AW is relatively low and the rc-mode occurs. In this case, the actor *chooses frame, script and action with the highest SEU* (102, equations 4.4, 4.5 and 4.6).

Next only *frame and script selections* are discussed in the text. It is argued that they (and not action selection) "usually follow the 'logic of appropriateness' (March and Olsen 1989). That is to say, there is a search for good reasons (Boudon 1996) in which actors aim at identifying the most appropriate alternative" (103).

Action selection "in the rc-mode is qualitatively different from" frame and script selection (103). For action "the actor typically will explicitly consider, evaluate, and weigh different and rather specific consequences" (103). Here "rational-choice theories are especially powerful" (103). We will return to this part of the MFS later. It is one of its most problematic parts. This is the assumption of "variable rationality."

Next the determinants of variable rationality are discussed (104 ff.). There are four determinants for the "mode of information processing" (104) that "the majority of dual-process theories agree on": high *opportunities*, strong *motivation*, low *effort* and low *accessibility* of cognitive elements make deliberation likely. The author explicitly refers to Fazio 1990. It is criticized that so far there is no formal model about these processes. Note that for each selection – frame, script or action selection – there is either an as- or a rc-mode chosen.

---

9 The author also writes that the activation of a frame is *determined* by the match. Because there is no separate definition of the AW, we assume that AW of a frame is *defined* by the match.

Again, SEU theory is applied. An equation for the SEU of the as-mode and one for the SEU of the rc-mode are formulated (105–106). The independent variables are the AWs of the alternatives (that is to say, frames, scripts or actions). We will not present the equations, but only the summary: "an actor selects (in) the rc-mode if, and only if, compared to an automatic – spontaneous selection, the additional utility of this mental activity exceeds its additional costs" (105).

3.6 The model of frame selection and the MODE model: Differences, similarities and their evaluation

This section focuses on a comparison of the MM and the MFS in several respects. After describing differences we discuss their plausibility. The MFS has come under attack by several authors (Etzrodt 2008; Lüdemann 1996; Lindenberg 2009; Opp 2010, 2017, 2019c). For limitations of space it is not possible to analyze this critique in detail. In this section, we will analyze those weaknesses which we consider most important.

(1) *The distinction between frames and scripts in the MFS is not part of the MM*. The question is whether this distinction is needed. If actions are to be explained it seems sufficient to identify only the factors that immediately influence the action. Why are other cognitions relevant which are stored in a wider mental model? The MM focuses on the cognitions directly relevant for the specific behavior. No other cognitive elements have direct effects on the behavior to be explained.

(2) *The assumption of the MFS that there is a causal (and hierarchical) sequence of frames, scripts and actions is not included in the MM*. The question is whether such a sequence is plausible. The MFS assumes that frames are activated first, then scripts and finally actions. Since scripts are included in frames, the sequence implies that individuals first activate a large class of cognitive elements that are not related for the actions. Only afterwards the action-related scripts are activated. The following example illustrates that this is implausible. Let a person want to cross a street and sees a traffic light. The individual will activate only information about what it means if the light is red or green, and which costs and benefits are to be expected if the street is crossed illegally. It is difficult to see why an individual will activate cognitive elements which are irrelevant in the specific situations. This effort can easily be saved.

(3) *The MFS assumes that entire frames or scripts are activated according to the "match" with the situation. The focus of the MM is on activation of single cognitive elements that are relevant for the behavior to be explained*. In the extreme case, an

entire situation may fit with all the cognitive elements. There is thus a perfect match. But in many other situations some elements might not fit. Assume a policeman in Germany at a traffic light wears a turban. There is thus one cue that does not fit to cognitions about policemen. This atypical element will probably be ignored because the actor reasons that everything else is what characterizes a policeman. This categorization will further activate expectations about the behavior of a policeman if someone crosses the street if the traffic light is red. This example illustrates that it is meaningful to concentrate on single cognitions that are relevant for performing an action and not on overall frames or scripts.

But if this is not accepted the question arises how such a general measure is constructed. This is particularly problematic if single cognitions are more or less important to an actor, as the policeman with a turban illustrates.

(4) *The MFS assumes that in the rc-mode, frame and script selection follow the "logic of appropriateness", whereas action selection is in accordance with rational choice theory. This is the assumption of "variable rationality." It is not included in the MM.* The first problem is that the rc- and as-mode are the extreme points of a quantitative variable, as advocates of the MFS themselves point out. For example, in choosing an apartment there will be most of the time some deliberation and some spontaneous decisions. When are those actions classified as being in the rc- or in the as-mode? In other words, where is the cutting point of the quantitative variable? As long as this is not clear, the MFS can actually not be applied in natural situations. It can only be decided ad hoc whether people are in the rc- or as-mode. But let us assume this problem is solved.

Another problem is the assumption that there is no subjective utility maximization in the as-mode. This is inconsistent with the application of SEU theory which is supposed to explain all three selections. Whatever the explanandum of the theory is, it implies that the alternative – be it a frame, script or action – with the highest SEU is chosen. This means that the actor engages in *subjective utility maximization*. This is clearly implied in the Kroneberg paper (2014) when the different costs and benefits of spontaneous and deliberate action are described (for example, 100–101).

"Variable rationality" means that there are two decision algorithms. Only one is subjective utility maximization. The other is acting "appropriately" or on "good reasons" (103). It is not clear what the difference is. Reference to Boudon's "good reasons" does not help because his "cognitive rationality" suffers from the same weakness as "variable rationality" (Opp 2014). The "logic of appropriateness" is as vague as acting on "good reasons."

In analyzing in detail the meanings of the different "rationalities" it is plausible that they all refer to enhancing the actors' well-being. Compare the following statements:

(1) P spends money for charity, because he thinks *this is appropriate*.
(2) P spends money for charity, because he thinks *there are good reasons for this*.

(3) P spends money for charity, because he thinks *this is best he can do in the present situation.*
(4) P spends money for charity, because he thinks *this maximizes his subjective utility*.

Each of these statements has the same meaning. The statements are different expressions of statement (4). For example, if P has good reasons for doing something, then this is subjectively the best course of action for him or her. Likewise, if P thinks that behavior B is "appropriate" (statement 1), this means that P does what in this situation is best for him.

The rejection of the assumption of subjective utility maximization in the MFS is *inconsistent with most of the social science literature*. This assumption is at least implicitly accepted by classical writers and used in numerous well confirmed social psychological theories (for details see Opp 2019a). If such a universally accepted hypothesis is rejected one would expect extensive evidence which is not provided.

The MM does not distinguish different "rationalities." As was said before, subjective utility maximization is assumed. According to the existing evidence, this seems to be a valid assumption.

(5) *The concept of "rationality" in the MFS needs clarification.* This concept is used in at least two meanings in the MFS. (1) "Rationality" refers to the extent to which people deliberate (for example, 100). "Variable" rationality means that people deliberate or do not deliberate.[10] (2) "Rationality" also seems to mean that people maximize utility (103). The question arises why the concept of rationality is useful at all (for a detailed discussion see Opp 2018). The MM does not contain the concept. This suggests that it is not needed.

(6) *The determinants of spontaneous or deliberate action are similar in the MFS and MM*. The variables opportunities, motivation, effort and accessibility are components of both models.

(7) *The MFS does not provide clear guidelines to identify the relevant cognitive elements of entire frames and scripts that explain behavior*. *The MM, in contrast, formulates hypotheses for single cognitive elements*. If there are no clear rules that allow to specify in detail the frames or scripts that are relevant for a behavior the application and test of the theory is ad hoc and allows an immunization of the theory (see Opp 2010). For example, what are the cognitive elements, that make up the frame or script for voting? The researcher might select those elements that have an effect

10 It is then strange if the author speaks of "rational" deliberation (100). It is not clear what a non-rational deliberation refers to.

and claim that this then confirms the MFS. It is not sufficient just to label frames as, for example, a "business frame" or "friendship frame" (117). These names suggest a clear definition, but what exactly the cognitive elements are these frames consist of is left open. In explaining reactions to punishment in an experiment, Esser (2018:17) introduces an altruism- and an egoism-frame. There is no specification of the cognitive elements these frames consist of.

(8) *The MFS does not explain misperception, in contrast to the MM*. The MFS addresses "selections" of frames, scripts and behavior. "Selection" means that something is chosen from a *given* set of objects. For example, from the situation a person is confronted with only certain elements or aspects are "selected" and more or less "matched" with existing cognitive elements. It is not clear whether "selection" also refers to biased perception. Even if this is meant it is not explained in the MFS when which misperception is to be expected. The MM includes such hypotheses, as has been said before.

(9) *Neither the MFS nor the MM explain the origins of beliefs in general and of preferences*. Although misperception is explained in the MM, it is difficult to see how in general the origin and change of beliefs can be explained. For example, how would the models explain when people accept certain conspiracy theories?

3.7 A critical comparison of the model of frame selection and rational choice theory

We saw that the basic assumptions of DPTs and the hypotheses of the MM in particular are an extension of the wide version of RCT. Only advocates of the MFS claim that parts of the model contradict RCT. For example, it is asserted that the MFS goes beyond "the economic dictum that behavior always follows incentives" (106). Furthermore, it is held that the MFS is the new overarching theory the social sciences have been looking for since their beginning (Esser 2017, see also Essser 2018). Among other things, the MFS allegedly explains when RCT fails (for example, Esser and Kroneberg 2010:84). In this section it is argued that the MFS is not contradictory to the wide version of RCT but only provides hypotheses that could extend RCT. It can thus not replace RCT.

3.7.1 Effects of the independent variables of the MFS

In order to determine whether the MFS contradicts RCT, a first step is to compare the independent variables of the two theories. The previous analysis shows that *all the variables of the MFS have an impact on incentives*. The framing of a situation and the activation of a script may lead to the activation of specific beliefs (which

may be wrong), preferences or of perceived behavioral alternatives. This is clearly implied if it is argued: "Akin to more recent developments in economics (. . .) one could specify *how actors' perceived choice set, preferences, and expectations vary depending on the selected frame and script* (103, italics added)." A clearer statement of the fact that the MFS extends RCT by specifying effects of its independent variables on independent variables of RCT is hardly possible. In a recent contribution Esser (2017) discusses findings of Tversky and Kahneman (1981) where different formulations ("framing") of identical propositions lead to different reactions of respondents. Esser tries to explain such effects by applying the MFS and argues that these results are the deathblow for "the" theory of rational action. But then he asserts that such framing effects (that is to say, effects of different formulations of the same statements) lead to a "neutralization of incentives" (511). In other words, framing effects change incentives, which, in turn, influence behavior.

3.7.2 Some applications of the model of frame selection

To determine whether the MFS and RCT are contradictory one could analyze applications of the MFS to specific explanatory problems (107–118). The question is to what extent these applications deal with changes of incentives. It is not possible for space limitations to discuss these applications in detail. Only a few notes must suffice.

(a) One hypothesis refers to the *effects of a strong "internalization" of a script* (107–108).[11] If a norm is relatively strong, individuals will not calculate, they follow the norm. Other incentives are thus not relevant in the present situation. This does not contradict a wide RCT because internalization of a norm is an incentive. RCT would assume that following a strong norm saves the costs of calculation and will thus be followed (for details see Opp 2017).

(b) The MFS has also been applied to explaining *voter turnout* (111–115). As is common in applications of the MFS, untested simplifying assumptions are made due to the lack of data. It is assumed that the "only relevant script is the civic duty norm" with a high accessibility of 1 which "clearly prescribes participation" (111). The internalized obligation to vote is a common variable in rational choice explanations of voting and not something new implied by the MFS.

According to the MFS it is further relevant for participating in an election that individuals "define the situation as 'election date'" (111). From a RCT perspective, one would predict that people who have a goal to vote have also an incentive to collect information about the *time* when an election takes place and about the *location* where

11 This is the first time that a script and not a norm is called "internalized."

one may cast his or her vote. Also in line with RCT is whether there is calculation or spontaneous voting. We refer to the previous discussion of applying VET for explaining participation. Note that "definition of the situation" activates incentives to go to the voting place. Furthermore, in explaining voting by any theory it is obviously assumed that people know the election date. Otherwise, people will not vote.

In explaining voting, it is first "necessary to identify the subset of measured incentives that have explanatory power" (113). This is exactly the procedure of RCT (for details see Opp 2001).

The major achievement of the MFS then are interaction effects of incentives "predicted by the MFS" (113). The idea is that "calculated incentives" interact with a civic duty measure: a strong voting norm reduces the impact of these incentives (that is to say, non-normative goals). RCT implies that actors with a strong norm do not want to bear the costs of calculation because following the norm is best for them anyway (Opp 2017). The interaction is thus also an implication of wide RCT.

It is striking that no effort is made to explore whether these interaction effects can be derived from a wide RCT as well (115, where apparently a narrow version of RCT is attacked). To conclude, the application of the MFS to explain voting is definitely not a demonstration of its superiority to RCT, it confirms the wide version of RCT.

(c) Kroneberg's discussion of *social movements and collective action* is suffering from major flaws. It is held that the MFS could provide a micro foundation for social movement research (116–117). The author notes similarities of the framing ideas of social movement theory and the MFS. The former has severe weaknesses that are not addressed (for details see Opp 2009). Again, a detailed analysis of the extent to which a wide version of RCT can be applied in social movement research is missing. There is further no analysis of the extent to which there are contradictions of the MFS and the wide RCT.

(d) The author provides a *list of applications of the MFS* (109–110, Table 4.1). If these examples are supposed to show the fruitfulness of the MFS, compared to other theories, one would expect a detailed discussion of alternative explanations. But this is missing. Even a cursory look at Table 4.1 indicates that each example specifies incentives that finally bring about the explanandum. Again, this list is not establishing any contradiction to RCT.

(e) Another example for the application of the MFS is Esser's analysis of research findings by Tversky and Kahneman (1981). Esser intends to show the failures of "the" theory of rational choice. The authors find that different formulations (that is to say, "framing") of tasks or choice situations, that are actually identical, lead to different reactions of the subjects in the experiment. In the following task (which is somewhat simplified) there is a population of 600 individuals and there is some disease. There are two programs to save people:

Program A –> 200 (from 600) people are saved – strong approval by a sample of respondents;

Program B –> 400 (from 600) people die – weak approval by a sample of respondents.

The puzzle is why the reactions to the two programs are different although the programs have the same effects: "200 from 600 people are saved" means the same as "400 from 600 people die." An answer could be that the experimental situation consists of different cues, in particular the words "saved" and "died." These seem to activate different cognitions. In explaining the reactions it needs to be specified which pre-existing cognitions led to which effects of the cues. What these cognitions are is to be determined empirically. One could conduct detailed in-depth interviews to find the reasoning of the subjects.

Whatever the reasoning of the subjects is: it influences incentives which, in turn affect the behavior, in this case utterance of an opinion. Strangely enough, this is what Esser himself asserts when he writes that the "framing effects" lead to the "neutralization of incentives" (511).

The relevance of incentives in this situation can be shown in the following way. The subjects in the experiment more or less approve the implementation of the programs. This means that they actually make a symbolic decision as if they had to implement the programs themselves. Applying SEU theory suggests that the decision depends on the perceived consequences of each program. What these consequences are should be determined empirically. It seems plausible that the subjects wish to save as many lives as possible. They might think erroneously that this consequence will be realized with program A to a higher extent than with program B. Thus, as Esser asserts, erroneous reasoning changes the incentives. But the incentives are relevant for the decision. The MFS variables are "causes of the causes."

There is a much easier – and more convincing – explanation of the different categorizations. In the MODE model, attitude or goal accessibility is a central variable, as was shown before. A strong accessibility determines the kind of categorization, if there are "multiple categorizable objects" (Fazio and Olson 2014: 157, where also supporting evidence is cited). The categorization is chosen that most closely resembles the accessible attitude. It is plausible that the attitude toward or the goal of "saving" lives is relatively strong. We will thus expect the findings reported in the work of Tversky and Kahneman. This is an explanation without applying the MFS.

3.8 Should dual-process theories always be applied?

Assume that DPTs consist of variables that influence incentives. Should DPTs then be applied in every explanation of behavior? The answer depends on the interests of the researcher. Most of the time scholars who apply RCT are only interested in the existing perceived incentives that directly determine action. This holds even if incentives may seem awkward from the perspective of the researcher. For example, in explaining why people commit crimes one would measure, among other things, beliefs of being punished. If it turns out that a group of individuals strongly underestimated punishment, researchers might be interested to explain misperceptions. Other researchers might be content with the measured beliefs and leave their explanation to further research. It is thus not meaningful to apply DPTs in every explanation.

Even one of the advocates of the MFS, Clemens Kroneberg, does not always apply the MFS in his work. An example is an article about Nation Building which presents a micro-macro model with an application of a wide version of RCT (Kroneberg and Wimmer 2012). One finds no mentioning of framing.

This implies that an application of RCT is not problematic if extensions are not addressed. It depends on the interest of the researcher whether, for example, biases or preferences are to be explained.

Accordingly, it is not to be criticized if most applications of a wide RCT focus on given incentives and not on their explanation. This holds, for example, for the literature on collective action in social movements research (Kittel and Opp 2018, with further references). Similarly, in criminology it is difficult enough to find the incentives that generate different kinds of crime. Therefore, most research in RCT concentrates on this task. There are, however, exceptions. Matsueda, Kreager, and Huizinga (2006) apply Bayesian learning hypotheses.

3.9 General conclusion

The conclusion from our discussion is that DPTs and the wide version of RCT are not contradictory, they complement each other. Figure 3.2 summarizes the relationship between the theories. DPTs are the *causes of causes* in the sense that they specify conditions that affect incentives for behavior. Note that this argument presupposes a particular version of RCT that seems most fruitful: it is based on all kinds of preferences of the actors and on their beliefs; it takes account of the limited cognitive capabilities of individuals and assumes that individuals try to reach, from their point of view, the best possible satisfaction when a decision is made.

One implication of the previous discussion is that DPTs confirm the fruitfulness of a wide version of RCT. In regard to the MODE model, it is clear that it is consistent

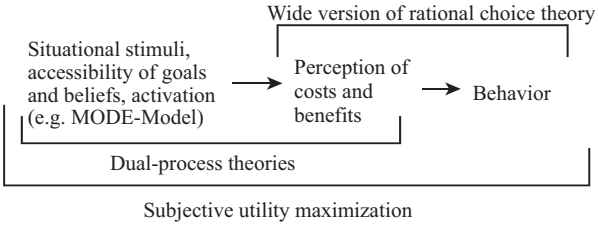


Figure 3.2: The relationship between dual-process theories and rational choice theory.

with RCT. *Advocates of the MFS actually show the fruitfulness of a wide version of RCT as well*, without being aware of it. SEU theory is applied for all explananda – see the equations referred to above. The introduction of "variable rationality" needs to be clarified and then empirically tested. This has not been done so far. At the present state of the discussion, the application of SEU for each of the MFS explananda and the denial of subjective utility maximization for the as-mode seem to be an *internal contradiction of theMFS*. Applying VET means that the action, frame and script with the highest SEU is chosen. This *means* that people choose what is best for them.

The model of Figure 3.2 shows that there are several explanatory stages. Variables of DPTs such as the MM and MFS consist of variables at the leftmost part of the model. These variables then affect incentives. They affect behavior. The variables at the left are a summary of several variables (see Figure 3.1). This part of the model could be broken down into different stages, as Figure 3.1 shows.

3.10 Further theory and research

"Behavior" in Figure 3.2 is usually understood as directly observable, external action. But it seems plausible to include also internal action such as thinking or focusing attention on some objects. These processes are governed by costs and benefits and subjective utility maximization as well, as has been said before. The application of SEU theory in the MFS suggests that such an extension of RCT or SEU theory would be useful. Systematic research is needed to test this claim.

Further research should also deal with integrating other social psychological theories with RCT and DPTs. Some of these theories such as dissonance or balance theory have the same or similar explananda as the MM and the MFS. Assume a person P likes O and suddenly learns that O has sympathies for terrorism, which P hates. P

will probably change his attitude toward O.[12] There should be a theoretical discussion of how to integrate DPTs with social psychological theories such as balance or dissonance theory which could even explain changes of preferences and beliefs.

As was said before, the MM originally includes attitudes as an independent variable. This was replaced by goals. It would be important to examine to what extent this replacement holds empirically.

A weakness of research in the tradition of the MFS is that it is based mainly on survey research. The MM is largely based on experimental studies. Perhaps experiments could supplement survey research to test hypotheses of the MFS.

An advantage of the MFS, compared to the MM, is that it has been applied to explain many *sociologically* interesting phenomena (for example, Kroneberg 2014: 109–110). As has been shown in our discussion of Kroneberg's voting study, often untested empirical assumptions referring to central propositions of the MFS are made. As long as these assumptions are not tested in a rigorous way, these studies can only be seen as exploratory research. A meta-analysis would be useful that explores which assumptions of the MFS have really been tested.

The MM, summarized in Figure 3.1, is actually a reconstruction. That is to say, it was sometimes not clear what exactly the causal relationships between the variables are. It is, in particular, important to disentangle the causal relationships between cues, pre-existing cognitions, accessibility, activation, selective perception (that is to say, incomplete and biased perception) and behavior. These relationships are not clear in the MFS either.

Advocates of the MFS emphasize as a strength its *formalization* which stands in contrast to other DPTs such as the MM. Although the formalization has been criticized (see Tutić 2015, 2016; Linnebach 2016) it clarifies the structure of a system of hypotheses and is therefore superior to a purely verbal formulation. However, precision of the structure of a theory is only one criterion of the quality of a theory. Other criteria are the precision of its concepts and its validity. As has been argued, many concepts are not clear, and rigorous tests without far-reaching untested assumptions are still missing.

What is the general conclusion from the previous analyses in regard to the question of what theory is to applied? In explaining behavior, the wide version of RCT still seems the best choice. If the researcher is interested in the causes of causes, the MM seems preferable. The previous analysis indicates that the MFS needs considerable improvement.

12 According to balance theory, there is a positive link from P to O, a negative link from O to X (O likes terrorism), and a negative link from P to X (P has a strong negative attitude toward terrorism). Balance theory would predict that P changes some attitude.

References

Balcetis, Emily, and David Dunning. 2006. "See What You Want to See: Motivational Influences on Visual Perception." *Journal of Personality and Social Psychology* 91(4):612–25.

Bargh, John A., Shelly Chaiken, Rajen Govender, and m Felicia Pratto. 1992. "The Generality of the Automatic Attitude Activation Effect." *Journal of Personality and Social Psychology* 62(6): 893–912.

Bargh, John A., and Melissa J. Ferguson. 2000. "Beyond Behaviorism: On the Automaticity of Higher Mental Processes." *Psychological Bulletin* 126(6):925–45.

Bargh, John A., Kay L. Schwader, Sareah E. Hailey, Rebecca L. Dyer, and Erica J. Boothby. 2012. "Automaticity in Social-Cognitive Processes." *Trends in Cognitive Sciences* 16(12):593–605.

Becker, Gary S. 1968. "Crime and Punishment. An Economic Approach." *Journal of Political Economy* 76(2):169–217.

Becker, Gary S. 1976. *The Economic Approach to Human Behavior*. Chicago and London: Chicago University Press.

Betsch, Tilmann, and Susanne Haberstroh (Eds.). 2005. *The Routines of Decision Making*. London: Lawrence Erlbaum.

Betsch, Tilmann, Susanne Haberstroh, and Cornelia Höhle. 2002. "Explaining Routinized Decision Making. A Review of Theories and Models." *Theory and Psychology* 12(4):453–88.

Bodenhausen, Galen V., and Andrew R. Todd. 2010. "Automatic Aspects of Judgment and Decision Making." Pp. 278–94 in *Handbook of Implicit Social Cognition. Measurement, Theory, and Applications*, edited by Bertram Gawronski and B. Keith Payne. New York and London: The Guilford Press.

Boudon, Raymond. 1996. "The 'Cognitivist Model.' A Generalized 'Rational-Choice-Model'." *Rationality and Society* 8(2):123–50.

Braun, Norman, and Thomas Gautschi. 2014. "Zwei Seelen wohnen, ach! in meiner Brust": Ein Rational-Choice-Modell innerer Konflikte." *Zeitschrift für Soziologie* 43(1):5–30.

Chaiken, Shelly, and Yaacov Trope. 1999. *Dual-Process Theories in Social Psychology*. New York and London: Guilford Press.

Cornish, Derek B, and Ronald V. Clarke. 2017 (2nd. ed). "The Rational Choice Perspective." Pp. 29–61 in *Environmental Criminology and Crime Analysis*, edited by Richard Wortley and Michael Townsley. Abington, UK: Routledge.

Deutsch, Roland, and Fritz Strack. 2010. "Building Blocks of Social Behavior Reflective and Impulsive Processes." Pp. 62–79 in *Handbook of Implicit Social Cognition. Measurement, Theory, and Applications*, edited by Bertram Gawronski and B. Keith Payne. New York and London: The Guilford Press.

Diekmann, Andreas, and Thomas Voss. 2004. "Die Theorie rationalen Handelns. Stand und Perspektiven." Pp. 13–32 in *Rational-Choice-Theorie in den Sozialwissenschaften*, edited by Andreas Diekman and Thomas Voss. München: R. Oldenbourg.

Elffers, Henk, and Jean-Louis van Gelder. 2017. "Criminal Decision Making: Time to Reject the Rational Choice Theory and Go on With the Dual Process Theory?" Pp. 124–31 in *Liber Amicorum Gerben Bruinsma*, edited by Catrien Bijleveld and Peter. van der Laan. Den Haag: Boom Juridische Uitgevers.

Esser, Hartmut. 1990. "'Habits', 'Frames' und 'Rational Choice'." *Zeitschrift für Soziologie* 19(4):231–47.

Esser, Hartmut. 2017. "When Prediction Fails. Reactions of Rational Choice Theory and Behavioral Economics to the Unexpected Appearance of Framing-Effects." Pp. 505–26 in *Social Dilemmas, Institutions and the Evolution of Cooperation*, edited by Ben Jann and Wojtek Przepiorka. New York: d.De Gruyter/Oldenbourg.

Esser, Hartmut. 2018. "Sanktionen, Reziprozität und die symbolische Konstruktion einer Kooperations-„Gemeinschaft"." *Zeitschrift für Soziologie* 47(1):8–28.

Esser, Hartmut, and Clemens Kroneberg. 2010. "Replik: Am besten nichts Neues?" Pp. 79–86 in *Sonderheft 50 der Kölner Zeitschrift für Soziologie und Sozialpsychologie: Soziologische Theorie kontrovers*, edited by Gert Albert and Steffen Sigmund. Wiesbaden: VS – Verlag für Sozialwissenschaften.

Esser, Hartmut, and Clemens Kroneberg. 2015. "An Integrative Theory of Action: The Model of Frame Selection." Pp. 63–85 in *Order on the Edge of Chaos: Social Psychology and the Problem of Social Order*, edited by Edward J. Lawler, R. Thye Shane, and Jeongkoo Yoon. Cambridge: Cambridge University Press.

Evans, Jonathan St. B. T. 2008. "Dual-Processing Accounts of Reasoning, Judgment, and Social Cognition." *Annual Review of Psychology* 59:255–78.

Evans, Jonathan St.B.T., and Keith Frankish (Eds.). 2009. *In Two Minds. Dual Processes and Beyond*. Oxford: Oxford University Press.

Fazio, Russell H. 1986. "How Do Attitudes Guide Behavior?" Pp. 204–43 in *Handbook of Motivation and Cognition. Foundations of Social Behavior*, edited by Richard M. Sorrentino and E. Tory Higgins. New York and London: Guilford Press.

Fazio, Russell H. 1990. "Multiple Processes by Which Attitudes Guide Behavior: The Mode Model as an Integrative Framework." Pp. 75–109 in *Advances in Experimental Social Psychology*, edited by Mark P. Zanna. San Diego: Academic Press.

Fazio, Russell H. 1995. "Attitudes as Object–Evaluation Associations: Determinants, Consequences, and Correlates of Attitude Accessibility." Pp. 247–82 in *Attitude Strength: Antecedents and Consequences*, edited by R.A. Petty and Jon A. Krosnick. Mahwah, NJ: Lawrence Erlbaum.

Fazio, Russell H. 2001. "On the Automatic Activation of Associated Evaluations: An Overview" *Cognition and Emotion* 15(2):115–41.

Fazio, Russell H. 2005. "Acting as We Feel: When and How Attitudes Guide Behavior." Pp. 41–62 in *Persuasion: Psychological Insights and Perspectives (2nd. ed.)*, edited by David R. Roskos-Ewoldsen and Timothy C. Brock. Thousand Oaks, CA: Sage.

Fazio, Russel H., and Michael A. Olson. 2014. "The MODE Model: Attitude-Behavior Processes as a Function of Motivation and Opportunity." Pp. 155–71 in *Dual-Process Theories of the Social Mind*, edited by Jeffrey W.S Sherman, Bertram Gawronski, and Yaacov Trope. New York: Guilford Press.

Fazio, Russel H., and Tamara Towles-Schwen. 1999. "The MODE Model of Attitude-Behavior Processes." Pp. 97–116 in *Dual Process Theories in Social Psychology*, edited by Shelly Chaiken and Yaacov Trope. New York: Guilford.

Feather, Norman T. (Ed.). 1982. *Expectations and Actions: Expectancy-Value Models in Psychology*. Hillsdale, N.J.: Lawrence Erlbaum.

Feather, Norman T. (Ed.) 1990. "Bridging the Gap betweeen Values and Actions. Recent Applications of the Expectancy-Value Model." Pp. 151–92 in *Handbook of Motivation and Cognition. Foundations of Social Behavior, volume 2*, edited by E.T Higgins and R.M. Sorrentino. New York: Guilford Press.

Fehr, Ernst, and Simon Gächter. 2000. "Cooperation and Punishment in Public Goods Experiments." *American Economic Review* 90(4):980–94.

Ferguson, Melissa J., and Shanette C. Porter. 2010. "What Is Implicit about Goal Pursuit?" Pp. 310–31 in *Handbook of Implicit Social Cognition. Measurement, Theory, and Applications*, edited by Bertram Gawronski and Shanette C. Payne. New York and London: Guilford Press.

Frankish, Keith, and Jonathan St.B.T. Evans. 2009. "The Duality of Mind: An Historical Perspective." Pp. 1–33 (not consecutively numbered) in *In Two Minds. Dual Processes and Beyond*, edited by Jonathan St.B.T. Evans and Keith Frankish. Oxford: Oxford University Press.

Gawronski, Bertram, and Laura A. Creighton. 2013. "Dual Process Theories." Pp. 282–312 in *The Oxford Handbook of Social Cognition*, edited by Donal E. Carlston. Oxford and New York: Oxford University Press.

Gilboa, Itzhak. 2010. *Rational Choice*. Cambridge, MA: MIT.

Hassin, Ran R., James S. Uleman, and John A. Bargh (Eds.). 2005. *The New Unconscious*. Oxford: Oxford University Press.

Hechter, Michael, Steven Pfaff, and Patrick Underwood. 2016. "Grievances and the Genesis of Rebellion: Mutiny in the Royal Navy, 1740 to 1820." *American Sociological Review* 81(1):165–89.

Hedström, Peter. 2005. *Dissecting the Social. On the Principles of Analytical Sociology*. Cambridge: Cambridge University Press.

Hedström, Peter, and Petri Ylikoski. 2014. "Analytical Sociology and Rational-Choice Theory." Pp. 57–70 in *Analytical Sociology. Actions and Networks*, edited by Gianluca Manzo. Chichester, UK: Wiley.

Houston, David A., and Russel H. Fazio. 1989. "Biased Processing as a Function of Attitude Accessibility: Making Objective Judgments Subjectively." *Social Cognition* 7(1):51–66.

Kahneman, Daniel. 2011. *Thinking. Fast and Slow*. London: Allen Lane.

Keren, Gideon, and Yaacov Schul. 2009. "Two Is Not Always Better Than One. A Critical Evaluation of Two-System Theories." *Perspectives on Psychological Science* 4(6):533–50.

Kittel, Bernhard, and Karl-Dieter Opp. 2019. "Dissecting the Conditions of Political Protest. An Exploration of Interaction Effects in the Explanation of Political Protest." *Sociological Inquiry* 89(1):67–93.

Kroneberg, Clemens. 2006. "The Definition of the Situation and Variable Rationality: The Model of Frame Seledtion as a General Theory of Action." Working paper *Sonderforschungsbereich 504, Universität Mannheim*.

Kroneberg, Clemens. 2014. "Frames, Scripts, and Variable Rationality: An Integrative Theory of Action." Pp. 97–123 in *Analytical Sociology: Actions and Networks*, edited by Gianluca Manzo. New York: Wiley.

Kroneberg, Clemens, and Andreas Wimmer. 2012. "Struggling over the Boundaries of Belonging: A Formal Model of Nation Building, Ethnic Closure, and Populism." *American Journal of Sociology* 118(1):176–230.

Kroneberg, Clemens, Meir Yaish, and Volker Stocké. 2010. "Norms and Rationality in Electoral Participation and in the Rescue of Jews in WWII: An Application of the Model of Frame Selection." *Rationality & Society* 22(1):3–36.

Kruglanski, Arie, Katarzyna Jasko, Marina Chernikova, Maxim Milyavsky, Maxim Babush, Conrad Baldner, and Antonio Pierro. 2015. "The Rocky Road From Attitudes to Behaviors: Charting the Goal Systemic Course of Actions." *Psychological Review advance online publication* http://dx.doi.org/10.1037/a0039541. 122.

Kruglanski, Arie W., and Tallie Freund. 1983. "The Freezing and Unfreezing of Lay-Inferences: Effects on Impressional Primacy, Ethnic Stereotyping, and Numerical Anchoring" *Journal of Experimental Psychology* 19(5):448–68.

Lindenberg, Siegwart. 2009. "Why Framing Should be All About the Impact of Goals on Cognitions and Evaluations." Pp. 53–79 in *Hartmut Essers Erklärende Soziologie. Kontroversen und Perspektiven*, edited by Paul Hill, Frank Kalter, Johannes Kopp, Clemens Kroneberg, and Rainer Schnell. Frankfurt and New York: Campus.

Linnebach, Patrick. 2016. "Erneut, warum eigentlich nicht? Replik zum Vorschlag, das Modell der Frame-Selektion zu axiomatisieren." *Zeitschrift für Soziologie* 45(2):122–35.

Lüdemann, Christian. 1996. "Der eindimensionale Akteur. Eine Kritik der Framing-Modelle von Siegwart Lindenberg und Hartmut Esser." *Zeitschrift für Soziologie* 25(4):278–88.

March, James G., and Johan P. Olsen. 1989. *Rediscovering Institutions. The Organizational Basis of Politics*. New York: Free Press.

Matsueda, Ross L., Derek A. Kreager, and Davis Huizinga. 2006. "Deterring Delinquents: A Rational Choice Model of Theft and Violence." *American Sociological Review* 71(1):95–122.

Mayerl, Jochen. 2009. *Kognitive Grundlagen sozialen Verhaltens. Framing, Einstellungen und Rationalität*. Wiesbaden: VS – Verlag für Sozialwissenschaften.

Mayerl, Jochen. 2010. "Die Low-Cost-Hypothese ist nicht genug. Eine Empirische Überprüfung von Varianten des Modells der Frame-Selektion zur besseren Vorhersage der Einflussstärke von Einstellungen auf Verhalten." *Zeitschrift für Soziologie* 39(1):38–59.

Olson, Michael A., and Russel H. Fazio. 2009. "Implicit and Explicit Measures of Attitudes. The Perspective of the MODE Model." Pp. 19–63 in *Attitudes. Insights from the New Implicit Measures*, edited by Richard E Petty, Russel H. Fazio, and Pablo Briñol. New York: Psychology Press.

Opp, Karl-Dieter. 1999. "Contending Conceptions of the Theory of Rational Action." *Journal of Theoretical Politics* 11(2):171–202.

Opp, Karl-Dieter. 2001. "Why Do People Vote? The Cognitive Illusion Proposition and Its Test." *Kyklos* 54(2/3):355–78.

Opp, Karl-Dieter. 2009. *Theories of Political Protest and Social Movements. A Multidisciplinary Introduction, Critique and Synthesis*. London and New York: Routledge

Opp, Karl-Dieter. 2010. "Frame-Selektion, Normen und Rationalität. Stärken und Schwächen des Modells der Frame-Selektion." Pp. 63–78 in *Sonderheft 50 der Kölner Zeitschrift für Soziologie und Sozialpsychologie: Soziologische Theorie kontrovers*, edited by Gert Albert and Steffen Sigmund. Wiesbaden: VS – Verlag für Sozialwissenschaften.

Opp, Karl-Dieter. 2013. "Norms and Rationality. Is Moral Behavior a Form of Rational Action?" *Theory & Decision* 74(3):383–409.

Opp, Karl-Dieter. 2014. "The Explanation of Everything. A Critical Assessment of Raymond Boudon's Theory Explaining Descriptive and Normative Beliefs, Attitudes, Preferences and Behavior." *Papers. Revista de Sociologia* 99(4):481–514.

Opp, Karl-Dieter. 2017. "When Do People Follow Norms and When Do They Pursue Their Interests? Implications of Dual-Process Models and Rational Choice Theory, Tested for Protest Participation." Pp. 119–41 in *Social Dilemmas, Institutions and the Evolution of Cooperation*, edited by Ben Jann and Wojtek Przepiorka. New York: de Gruyter/Oldenbourg.

Opp, Karl-Dieter. 2018. "Do the Social Sciences Need the Concept of "Rationality"? Notes on the Obsession with a Concept." Pp. 191–217 in *The Mystery of Rationality. Mind, Beliefs and the Social Sciences*, edited by Francesco Di Iorio and Gérald Bronner. Wiesbaden: VS Springer.

Opp, Karl-Dieter. 2019a. "Are Individuals Utility Maximizers? Empirical Evidence and Possible Alternative Decision Algorithms." in *Grundlagen – Methoden – Anwendungen in den Sozialwissenschaften*. Pp. 421–440 in *Festschrift für Steffen M. Kühnel*, edited by Methodenzentrum Sozialwissenschaften Universität Göttingen. New York: Springer VS.

Opp, Karl-Dieter. 2019b. "Rational Choice Theory and Methodological Individualism." in *The Cambridge Handbook of Social Theory*, edited by Peter Kivisto. Cambridge: Cambridge University Press. Forthcoming.

Opp, Karl-Dieter. 2019c. "Die Theorie rationalen Handelns, das Modell der Frame-Selektion und die Wirkungen von Bestrafungen auf Kooperation. Eine Diskussion von Hartmut Essers Erklärung der Ergebnisse eines Experiments von Fehr und Gächter (2000, 2002)." *Zeitschrift für Soziologie* 48(2):97–115.

Opp, Karl-Dieter. 2020. *Analytical Criminology. Integrating Explanations of Crime and Deviant Behavior*. London and New York: Routledge.

Riker, William H., and Peter C. Ordeshook. 1968. "A Theory of the Calculus of Voting." *American Political Science Review* 65:25–42.

Riker, William H., and Peter C. Ordeshook. 1973. *An Introduction to Positive Political Theory*. Englewood Cliffs, N.J.: Prentice Hall.

Sandler, Todd. 2001. *Economic Concepts for the Social Sciences*. Cambridge: Cambridge University Press.

Sherman, Jeffrey W., Bertram Gawronski, and Yaacov Trope (Eds.). 2014. *Dual-Process Theories of the Social Mind*. New York: Guilford Press.

Srull, Thomas K., and Robert S. Wyer jr. 1986. "The Role of Chronic and Temporary Goals in Social Information Processing." Pp. 503–49 in *Handbook of Motivation and Cognition. Foundations of Social Behavior*, edited by Richard M. Sorrentino and E. Tory Higgins. New York and London: The Guilford Press.

Stigler, George J. 1950a. "The Development of Utility Theory. I." *Journal of Political Economy* 58(4): 307–27.

Stigler, George J. 1950b. "The Development of Utility Theory. II." *Journal of Political Economy* 58(5): 373–96.

Thomas, Kyle J., and Jean Marie McGloin. 2013. "A Dual-Systems Approach for Understanding Differential Susceptibility to Processes of Peer Influence." *Criminology* 51(2):435–74.

Tutić, Andreas. 2015. "Warum denn eigentlich nicht? Zur Axiomatisierung soziologischer Handlungstheorie." *Zeitschrift für Soziologie* 44(2):83–98.

Tutić, Andreas. 2016. "Zur Interpretation entscheidungstheoretischer Kalküle – Eine Erwiderung." *Zeitschrift für Soziologie* 45(2):136–44.

Tutić, Andreas, Thomas Voss, and Ulf Liebe. 2017. "Low-Cost-Hypothese und Rationalität. Eine neue theoretische Herleitung und einige Implikationen." *Kölner Zeitschrift für Soziologie und Sozialpsychologie* 69(4):651–72.

Tversky, Amos, and Daniel Kahneman. 1974. "Judgment under Uncertainty: Heuristics and Biases. Biases in Judgments Reveal Some Heuristics of Thinking Under Uncertainty." *Science* 185:1124–31.

Tversky, Amos, and Daniel Kahneman. 1981. "The Framing of Decisions and the Psychology of Choice." *Science* 211:453–58.

Vazsonyi, Alexander T., and Albert J. Ksinan. 2017. "Understanding Deviance through the Dual Systems Model: Converging Evidence for Criminology and Developmental Sciences." *Personality and Individual Differences* 111(1):58–64.

Wilson, Timothy D. 2002. *Strangers to Ourselves. Discovering the Adaptive Unconscious*. Cambridge, Mass.: Belknap Press.

Marcel A.L.M. van Assen and Jacob Dijkstra

4 Too Simple Models in Sociology: The Case of Exchange

Abstract: When examining a social phenomenon, theoretical and empirical sociologists require a model. Although models are by definition simplified representations of theories of reality, sociologists typically argue that the micro-level model of individual behavior should be simplified, but not the model of the macro-level system (including macro-micro and micro-macro links) including the social interactions. Using the example of research on exchange we theoretically and empirically demonstrate the possibly disastrous consequences of overly simplifying the model of the macro-level system. We show that by oversimplification, the mainstream model of exchange (so-called split pool exchange) precludes explaining the macro-level phenomenon. Consequently, we question the ecological validity of exchange research for real-life exchanges. We conclude by listing advantages of a more complex and realistic macro-level model of exchange (that is, pure exchange) for research in the social sciences.

> Everything should be made as simple as possible, but not simpler
>
> (saying attributed to Albert Einstein)

4.1 Introduction

A central problem in sociology is that of accounting for the functioning of a social system (Coleman 1990). Coleman (for example, 1987, 1990) proposed a scheme for explaining macro-level relationships. This scheme contains three links, relating the macro-level (that isthat is, the level of social phenomena) to the micro-level (that is, the level of individuals): (i) the macro-micro link, representing how social conditions affect (assumptions about or a theory of) conditions of individuals, (ii) the micro-level model or theory transforming micro-conditions into micro-outcomes, and (iii) the micro-macro link transforming the micro-outcomes into the macro-outcomes of interest. Discussions of research in sociology focus on the shape and detail of each of those links in sociological theory when attempting to explain macro-level phenomena. For instance, Coleman (1987) argues that sociologists traditionally

Marcel A.L.M. van Assen, Department of Sociology, Faculty of Social and Behavioural Sciences, Utrecht University, the Netherlands; Department of Methodology and Statistics, Tilburg School of Social and Behavioral Sciences, Tilburg University, the Netherlands
Jacob Dijkstra, Department of Sociology, Faculty of Social and Behavioural Sciences, University of Groningen, the Netherlands

https://doi.org/10.1515/9783110647495-004

often use aggregation to transform individual outcomes to macro-level outcomes, and convincingly shows that aggregation often does not qualify as a valid model for this transformation.

A related fundamental discussion involves the detail and accuracy of the micro-micro link or model of individual behavior. Whereas Coleman and many other rational choice sociologists proposed to keep the micro-level model as simple as possible and not necessarily realistic or accurate (see also Friedman, 1953, for a defense of this position), others, including analytical sociologists such as Hedström (2005: 34–38), contend that the micro-level model should be psychologically and sociologically realistic. Without going into the specific details of this discussion, Hedström (2005: 35) argues that ("within the restrictions imposed by psychological and sociological plausibility") "we should seek a theory that is as simple as possible". Counterbalancing this KISS[1] principle is the "but not too simple" addition in the motto of this paper, adequately formulated in Lindenberg's (2001) principle of *sufficient complexity*. This principle entails that model assumptions should be realistic enough to allow a description and explanation of the social phenomenon, without "assuming away" essential ingredients of that phenomenon.

In this paper we examine the (in)sufficient complexity of macro-level models of the social phenomenon to be explained. We believe the issue of sufficiently complex macro-level models has received too little attention in the literature. With 'macro-level models' we mean the model of the social situation including social interactions, and the macro-micro and micro-macro links, but excluding elements that solely reside at the micro-level such as the model of actor preferences and actor behavior. We demonstrate, using the case of exchange research, how simplifying the macro-level model may preclude explaining the social phenomenon, by assuming away its essential ingredients. More specifically, we first demonstrate that the *split pool exchange* task that is commonly used in research on exchange in economics, social psychology, and sociology, assumes away essential characteristics of typical real-world exchange. In the split pool exchange task actors split a constant common resource pool, whereas in typical real-world exchanges actors transfer units of goods they value differently (called *pure exchange*). As such, the simplification of the macro-level in our demonstration concerns the *task* or social interaction between individuals, and not the social structure (for instance. the network) in which these interactions are embedded.

We provide two empirical demonstrations of the consequences of simplifying the model of exchange. The first demonstration concerns bilateral exchange and shows that outcomes of the split pool exchange task are very different from outcomes of pure exchange (Dijkstra and Van Assen 2008). The second demonstration, concerning exchange in larger social structures such as groups by Dogan and Van Assen (2009), additionally suggests that alternative types of profitable exchange

1 Originally referring to "Keep it simple, stupid"; see "KISS principle" (2019).

opportunities exist that differ in how easily they are detected by actors. We believe this differential detectability to be an essential element of real-life exchange, and show that the split pool exchange task assumes it away. We conclude with listing other unfortunate consequences of simplifying pure exchange to the split pool task when investigating exchange and related phenomena. Based on our theoretical analysis and empirical demonstrations we argue that the ecological validity of exchange research employing the split pool exchange task as a stand-in for real-life exchange is questionable at best. By briefly discussing another example, the case of the Prisoner's Dilemma game as a standard model of social dilemmas, we hope to convince the reader that using too simplified macro-level models in the social sciences is a more general issue. We therefore hope that our paper will increase sociologists' sensitivity to the applicability of the principle of sufficient complexity to the macro level (and not just to the micro level), when modeling social phenomena.

4.2 Exchange

Exchange is intensively studied in economics as well as sociology and social psychology. An exchange situation can be defined as a situation involving actors who have the opportunity to collaborate for the benefits of all actors involved, which is similar to Nash's (1950: 155) definition of a bargaining situation. While bargaining and exchange became the object of research of economists in the 19th century (Edgeworth 1881), exchange was only beginning to be studied in social psychology and sociology in the 1950s. For instance, Homans (1958: 606) stated that "social behavior is an exchange of goods, material goods but also non-material ones, such as the symbols of approval and prestige".

The prototypical example of an exchange situation is *pure exchange*, with actors holding bundles of goods they wish to exchange for other goods. We refer to the goods held at the outset as *initial endowments*. Based on the preferences or utilities of actors for these goods, actors may exchange their initial endowments to obtain their *final endowments*. An example of a bilateral pure exchange situation is where actor A has 18 units of good X and actor B has 30 units of good Y, and A values X and Y equally, whereas B values each unit of X five times as much as a unit of Y. See the last rows of Table 4.1 (Condition 5) for a representation of this exchange situation in numbers. In economics two main approaches may be distinguished for predicting the final endowments after exchange (for instance, Hildenbrand and Kirman 1998); the general equilibrium approach and Edgeworth's (1881) approach using the *core*. The core is that set of possible final endowments that cannot be improved upon by any coalition of actors (Hildenbrand and Kirman 1998: 16). In the case of our bilateral exchange example the only coalitions of actors to consider are the {A, B} dyad and the {A} and {B} singletons.

Table 4.1: Endowments (E) and utilities (U) of goods X and Y for the pure exchange conditions in Dijkstra and Van Assen (2008).

Condition	Actors	A		B	
	Goods	X	Y	X	Y
2	E	1	0	0	90
	U	18	1	90	1
3	E	18	0	0	90
	U	1	1	5	1
4	E	18	0	0	30
	U	3	3	5	1
5 (typical pure exchange; Fig.1)	E	18	0	0	30
	U	1	1	5	1

Source: Adapted from Dijkstra and Van Assen (2008: 27)

Figure 4.1 represents the payoff space of all outcomes of the example of bilateral pure exchange, with the shaded area representing all mutually beneficial payoffs for A and B, and the "frontier" representing the core or the set of Pareto-optimal payoffs. The payoffs (U(a), U(b)) = (0, 72) are obtained if A transfers all of his 18 X to B in

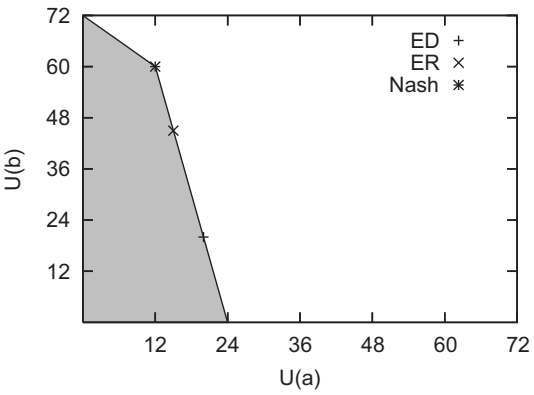


Figure 4.1: Payoff space of the bilateral pure exchange example of condition 5 of Dijkstra and Van Assen (2008).
Note: U(a) and U(b) represent actor A's and B's utility gain after the exchange. ED and ER indicate equidependence and equiresistance, respectively.
Source: Reprinted from Dijkstra and Van Assen (2008: 20)

exchange for 18 of B's Y, yielding $U(a) = 18 \times 1 - 18 \times 1 = 0$ and $U(b) = 18 \times 5 - 18 \times 1 = 72$. The payoffs (U(a), U(b)) = (12, 60) are obtained if A still transfers all of his X, but B now also transfers all of his 30Y, yielding $U(a) = 30 \times 1 - 18 \times 1 = 12$ and $U(b) = 18 \times 5 - 30 \times 1 = 60$. The frontier connecting these two payoffs results from A transferring all his X to B and B transferring from 18 to 30 of his Y to A. Finally, payoffs (U(a), U(b)) = (24, 0) result from B transferring all his 30 Y to A, with A compensating this by transferring 6 of his X, yielding $U(a) = 30 \times 1 - 6 \times 1 = 24$ and $U(b) = 6 \times 5 - 30 \times 1 = 0$. Hence, the payoff frontier connecting the last payoffs is obtained by exchanges where B transfers all of his Y in exchange for 6 to 18 of A's X.

In most experimental studies on exchange in social psychology and sociology (see *Social Networks*, *14*, and Van Assen, 2003, for an overview), and in behavioral economics (for instance, Camerer 2003; Roth 1995), an abstraction of pure exchange is used in which exchange is conceptualized as the opportunity of two actors to split a resource pool, rather than to exchange actual units of goods. The intuition underlying this *split pool exchange* is that since exchange creates surplus value for both exchange partners, it might as well be represented by negotiations over that surplus value, rather than over the transfer of actual goods. The actors' task in split pool exchange is thus to negotiate over the split of a pool of valuable points, typically 24, which have the same value for both exchange partners. If the two actors agree on the division, the points are divided according to the agreement, whereas they do not obtain any points if they fail to reach agreement. Importantly, the entire pool of points must be divided in the agreement. Figure 4.2 represents the payoff space of split pool exchange with a pool of 72 points.

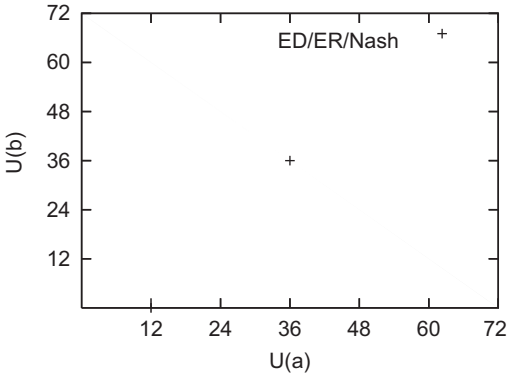


Figure 4.2: Payoff space of abstraction of bilateral pure exchange (split pool exchange) of condition 1 of Dijkstra and Van Assen (2008).
Notes: U(a) and U(b) represent actor A's and B's utility gain after the exchange. ED and ER indicate equidependence and equiresistance, respectively.
Source: Reprinted form Dijkstra and Van Assen (2008: 19)

The use of split pool exchange to study exchange has commonly been justified by stating (or often implicitly assuming) that split pool exchange is equivalent to pure exchange (see Van Assen 2001, and Dijkstra and Van Assen 2008, for references and citations). However, some doubts have been raised concerning the validity of this equivalence statement. For instance, Bonachich (1992: 22) noted that "nothing is actually exchanged in these experiments." Indeed, pure exchange and split pool exchange are different in four respects (Dijkstra and Van Assen 2008: 21):

(i) The *task*; in pure exchange actors exchange resources, whereas in split pool exchange actors split a fixed pool of points.
(ii) *Pareto efficiency*; in split pool exchange Pareto efficiency is enforced because the entire pool must be divided, whereas Pareto efficiency is not guaranteed in pure exchange, as indicated by the shaded area in Figure 4.2.
(iii) *Constant-sum*; split pool exchange entails a constant sum of payoffs of both actors, which is typically not true for pure exchange.
(iv) *Equal maximum*; in split pool exchange both actors' maximum payoffs are equal, which is generally not true in pure exchange.

In other words, aside from being two different tasks, split pool exchange is less complex than pure bilateral exchange; whereas split pool exchange can be represented by one parameter (the size of the pool), bilateral exchange needs at least eight parameters (for each actor, two parameters for initial endowments X and Y and two more for the utilities of X and Y).

Note how representing pure exchange by split pool exchange entails a simplification of the macro-model of social interaction. The fundamental question is, whether this simplification yields outcomes and conclusions that can still be generalized to pure exchange. We argue that this is not the case and that simplifying pure exchange to split pool exchange violates the principle of sufficient complexity (Lindenberg 2001), at the macro level. Thus, we argue that the assumptions underlying split pool exchange are not realistic enough to allow a description of the phenomenon (that is, pure exchange) to be explained. More specifically, using split pool exchange one cannot explain possible inefficiency of exchange outcomes as inefficiency is "assumed away", and one cannot explain how actors may deal with inequality of exchange opportunities as it assumes constant-sum splits with equal maxima to both actors.

These differences between split pool exchange and pure exchange have substantial consequences for theory and predicted outcomes. Van Assen (2001: Chapter 7) shows that well-known solutions to bilateral bargaining, such as Nash's bargaining solution (Nash 1950), the Raiffa-Kalai-Smorodinsky (RKS) solution (Kalai and Smorodinsky 1975) and the kernel solution (Shubik 1982) yield identical predictions for split pool exchanges; the 12–12 split. These solutions are shown in Figure 4.2 using the labels equidependence (corresponding to the kernel) and equiresistance (corresponding to the RKS) more familiar in the sociology literature. These solution

concepts underlie two prominent theories of network exchange known as power-dependence theory (Cook and Emerson 1978) and network exchange theory (for instance, Willer 1999), respectively. Notwithstanding their equality in split pool exchange, these solutions are typically (and sometimes very) different in pure exchange, as is illustrated in Figure 4.1. Consequently, split pool exchange precludes testing basic principles of bargaining (as formalized in the alternative solutions) in exchange. Finally, the change of task from a complex to a simple one may itself also affect the outcomes of bargaining. Because of all the differences and their implications, we contend that outcomes and conclusions of research using split pool exchange are not ecologically valid for pure exchange.

4.3 Empirical demonstration: Bilateral exchange

A comparison of Figures 4.1 and 4.2 demonstrates that pure exchange and split pool exchange are not equivalent in terms of the payoff possibilities. The remaining empirical question then concerns the extent to which results and conclusions of split pool exchange studies can be generalized to pure exchange. Dijkstra and Van Assen (2008) set out to answer this fundamental question by comparing experimental bilateral bargaining outcomes in split pool exchange and pure exchange situations.

Their study contains five conditions. The extremes, Condition 1 and Condition 5, are a bilateral split pool exchange with 72 points (Figure 4.2 shows its payoff space), and the bilateral pure exchange introduced in our example above (see last two rows of Table 4.1, and Figure 4.1), respectively. These two conditions differ in all four respects listed above: experimental task, enforced Pareto efficiency, constant-sum payoffs, and equal maxima for the exchange partners. The remaining 3 conditions systematically fill the interval between Condition 1 and Condition 5. Condition 2 is a pure exchange situation that differs from Condition 1 only in experimental task. Thus, Pareto efficiency is enforced in Condition 2, the sum of payoffs of both partners is 72 for any agreement, and the maxima for both partners are 72 (that is, what a partner can maximally obtain when at the same time the other partner gains nothing). Condition 3 is a pure exchange situation that allows Pareto inefficient agreements, but is otherwise identical to Condition 2. Thus, in Condition 3 the sum of payoffs in all Pareto efficient agreement is 72, and so are the maxima of both partners. In the pure exchange situation of Condition 4, payoffs of Pareto efficient agreements do *not* all sum to the same constant, but otherwise Condition 4 is identical to Condition 3. Particularly, the maximum payoff is 72 for both exchange partners. This latter characteristic is finally abandoned in Condition 5, which has maximum payoffs of 24 and 72 for A and B, respectively. The initial endowments and utilities for conditions 2 through 5 are given in Table 4.1.

In the experiment reported by Dijkstra and Van Assen (2008) 124 subjects participated. The design was between-subjects, so each subject was active in only one of the conditions. Subjects exchanged for at most six rounds, and did so with the same partner throughout. Dijkstra and Van Assen (2008) used a full information design, meaning subjects had all information about their own and each other's endowments and payoffs. After each round the exchange situation was reset (the pool was replenished in Condition 1, and all subjects were given new initial endowments in the other conditions). Points earned in the experiment were converted to money and paid at the end. The ExNet program developed by David Willer and colleagues at University of South Carolina was used to run the experiments. See Dijkstra and Van Assen (2008) for more details on design and procedure.

Dijkstra and Van Assen (2008) developed hypotheses on a number of outcome variables, most notably average payoffs, variance in payoffs, and Pareto efficiency of agreements. The first main result is that three prominent existing bargaining theories (the Nash bargaining solution (1950), the Raiffa-Kalai-Smorodinsky (1975) solution and the Kernel solution (Shubik 1982)) predict average payoffs accurately only in constant-sum bilateral exchange (conditions 1 through 3). Given agreement, the proportion of equal payoff agreements (predicted by the Kernel solution) is considerable, ranging from a low of 0.41 in Condition 5 to a high of 0.67 in Condition 3. The second main result is that variances in payoffs are much larger in pure exchange than in split pool exchange. Tellingly, the payoff variance in Condition 5 is over 50 times larger (!) than the variance in Condition 1. The final main result is that many agreements, when reached, are Pareto inefficient, from a low of 27% in Condition 3 to a high of 61% in Condition 5. In this respect there is evidence of learning: across the rounds of exchange, the probability of Pareto efficiency is increasing.

These results show that the simplifications implicit in the move from pure exchange to split pool exchange 'abstract away' crucial aspects of the phenomenon to be explained; bargaining behavior in (bilateral) exchange. Such bargaining behavior in the real world very often involves the transfer of actual goods, that is, is best conceptualized as pure exchange. The results of Dijkstra and Van Assen (2008) show that such bargaining behavior (i) is much less accurately predicted by existing bargaining theories than suggested by split pool exchange studies, (ii) seems much less predictable (much more variable) than suggested by split pool exchange, in the first place, and (iii) much more frequently leads to Pareto inefficient outcomes than suggested by split pool exchange studies. Note how the second implication (related to the higher variability of outcomes in pure exchange) points to the need of theoretically including covariates (such as bargaining skills, or mental models of the bargaining situation) that have hitherto been neglected in many bargaining and exchange theories. This drives home the point that accurate macro-models are crucial for fruitful theory development.

The results also carry implications for network exchange theory. The vast majority of studies in this field have employed the split pool exchange paradigm. The results found by Dijkstra and Van Assen (2008) strongly suggest that the relative accuracy of theoretical predictions in exchange networks might be an artifact of just this paradigm. How would theoretical predictions of existing network exchange theories fare in networks of pure exchange? We currently do not know.

Finally, some theories and concepts of power in social networks (such as Burt's concept of brokerage, or Bonacich centrality) rely on notions of competitive advantage of nodes due to their positions in the network. When motivating the concept of brokerage, Burt (2005: 39) explicitly refers to the findings from experimental network exchange research. Research that has consistently used an overly simplified macro-model of the situation it tries to understand, thereby jeopardizing the validity of its claims.

4.4 Empirical demonstration: Exchange in groups

Most research on exchange in sociology focuses on network exchange, more specifically the effect of network structure (where ties represent restrictions on who can exchange with whom) on exchange outcomes of actors occupying nodes in the network. Most of network exchange research uses the split pool exchange task, with both actors in each tie of the network having the opportunity to split a resource pool of commonly 24 points, implying equally valuable exchange relations across the network. Usually, each actor has the restriction to exchange only once (for instance, see Van Assen, 2003, for an overview of this network exchange research in sociology). Just like outcomes of bilateral exchange are different from those of split pool exchange, one may wonder if exchange in groups modelled with split pool exchange brings in additional violations of sufficient complexity. That is, does split pool exchange 'assume away' ingredients of exchange in groups that are present in pure exchange in groups?

Dogan and Van Assen (2009) reanalyzed the experimental data of Michener, Cohen, and Sørensen (1975, 1977) on pure exchange in small groups. Michener et al. (1975) and Michener et al. (1977) examined pure exchange situations with three actors exchanging four goods, and with four actors exchanging five goods, respectively. Table 4.2 presents an example of a pure exchange situation used in Michener et al. (1975). At the start of the experimental session on that exchange situation three actors (Alpha, Beta, Gamma) received initial endowments (to the right of the table) of four goods (Red, White, Blue, Yellow) in multiples of 10 units. Additionally, they were told what their payoffs (utilities; to the left of the table) would be a function of their final endowments. Payoffs were linear in endowments, such that Alpha's coefficient '0.8' for Red means that at the end of the experimental session Alpha obtains a payoff of 0.8 monetary units for each additional unit of Red he has.

Table 4.2: One pure exchange situation used in the experiment of Michener et al. (1975).

	Utilities				Endowments		
	Alpha	**Beta**	**Gamma**		**Alpha**	**Beta**	**Gamma**
Red	0.8	0.6	0.8	Red	0	90	10
White	0.0	0.1	0.0	White	40	10	50
Blue	0.0	0.2	0.2	Blue	10	20	70
Yellow	0.2	0.1	0.0	Yellow	70	10	20

Note: Names of colors and names of Greek characters represent endowments and actors, respectively.

Each of Michener et al.'s studies examined three different pure exchange situations, with 16 groups of three persons and 12 groups of four persons for each configuration, totaling 144 participants in each study. The experimental procedure was identical in all exchange situations (see Michener et al. 1975, 1977, for details). Participants were seated around the table and had complete information on all subjects' endowments and utilities. Only bilateral deals were allowed, and participants were free to talk during the experiment. An experimental session stopped if participants decided they no longer wanted to trade. Participants' actual payoffs were determined by a show-up fee and their final endowments multiplied by the utilities of their endowments.

When (re)analyzing the pure exchange situations of Michener et al. (1975, 1977) two facts are observed almost instantly. First, potential profits vary wildly both within and across exchange opportunities, in contrast with the constant-sum payoffs of exchange opportunities using split pool exchange. Equally profitable exchange opportunities can be expected to be exceedingly rare in pure exchange situations (Van Assen 2001). Second, some profitable exchange opportunities seem easier to detect than others. Dogan and Van Assen (2009) hypothesized exchange opportunities that are more difficult to detect are less likely to be carried out. They distinguished three heuristics to detect profitable bilateral exchange opportunities. Increasing in sophistication, these heuristics are:

(i) Exchange a good you do not want yourself for some good you value ('zero').
(ii) Exchange a good you want for some good you like more ('absolute').
(iii) Exchange a good you want for some good you like relatively more ('relative').

That is, actor A may want to exchange good X for good Y of B because a unit of Y is more valuable to A than a unit of X (heuristic (ii)), or because the ratio (value of unit Y / value of unit X) is higher for A than for B (heuristic (iii)). For instance, if this ratio is ½ for A and ¼ for B, an exchange is beneficial to both actors if A gives up some of his most preferred good X in exchange for 2 to 4 times as much of good Y.

On the basis of these three heuristics they identified four different exchange opportunities:

D0: Actors have *d*ifferent preference orders; at least one has *0* interest in one good
S0: Actors have the *s*ame preference order, and one actor has *0* interest in one good
Db: Actors have *d*ifferent preference orders; *b*oth actors have interest in both goods
Sb: Actors have the *s*ame preference order; *b*oth actors have interest in both goods

They hypothesize that D0 and S0 are particularly easy to detect, as actors can be expected to willingly give up a good they do not value in exchange for some other good, and expect that Sb is particularly difficult to detect as both actors value the same good the most.

Dogan and Van Assen (2009) tested their hypotheses by comparing the potential exchange opportunities at the onset of the experimental sessions to those after the sessions ended, based on the final endowments of the actors. The potential exchange opportunities at the outset were derived from tables like Table 4.2. The opportunities at the end were derived from the table with final endowments after the exchanges (table(s) not shown here, see Dogan and Van Assen). Dogan and Van Assen computed both the frequencies of opportunities (D0, S0, Db, Sb) as well as the potential gains from these opportunities, for both the initial and final endowment tables. Both the frequencies and potential gains were computed with a bilateral exchange model assuming that both actors obtain an equal gain in their exchange, which seems a reasonable assumption given the outcomes of Dijkstra and Van Assen (2008) discussed previously. For instance, a frequency of 5 for Db means that five mutually beneficial bilateral exchanges could be carried out of type Db (with actors having different preference orders for both goods and having interest in both goods) given the endowments, The 'potential gain' simply sums the gains of these five opportunities, assuming the bilateral exchange model is used to determine the rate of exchange.

If actors are fully rational, actors should have no problem to detect exchange opportunities of any type, including exchanges of type Sb. Consequently, no mutually beneficial exchange opportunities would remain at the end of exchanging. However, if some opportunities are more difficult to detect than others, we would expect that (i) particularly 'difficult' exchange opportunities remain at the end of the experiment, and (ii) the potential gain of 'difficult' exchange opportunities at the end of the experiment is higher than for less 'difficult' opportunities. Dogan and Van Assen's predictions were confirmed. At the outset the frequency of opportunities (D0, S0, Db, Sb) was (57, 27, 2, 2), and at the end (2, 9, 2, 7), meaning that both the relative and absolute number of complex exchange opportunities increased.[2] Similarly, comparing the

2 Three different statistical tests were conducted on the frequency of opportunities, all significant at 0.01. The statistical test comparing potential gains of D0 and S0 on the one hand to Db and Sb on the other hand resulted in $p = 0.004$.

potential gains at the start (460.86, 107.9, 40, 3.5) to those at the end (0.17, 0.007, 1.17, 3.96) clearly demonstrated that actors did fully realize the potential of simpler opportunities, but not of the complex ones, particularly Sb.

To conclude, Dogan and Van Assen (2009) convincingly showed that potential exchange opportunities may be very different in two crucial respects. Some are naturally *more valuable* than others, and importantly, some may be *more difficult to detect* than others. Particularly profitable exchange opportunities where both actors like the same good most were difficult to detect, suggesting that actors were reluctant to give up the good that they liked most, thereby violating the fully rational aforementioned 'relative' heuristic (iii). As split pool exchanges 'assumes away' these phenomena, it provides an additional violation of the principle of sufficient complexity in the context of group exchange. Note that the difficulty to detect an exchange opportunity may also interact with the effect of the network structure on the outcomes of exchange. For example, consider an exchange network with one so-called powerful actor having many opportunities of a complex type, who can exclude others from exchanging. If other more peripheral actors in the network who also have opportunities of a simple type, the central actor may not be able to realize his full potential.

4.5 Other advantages of sufficiently complex macro-level models

Overly simplifying the model of the social phenomenon, or how it translates the macro-conditions into the micro-conditions not only may have consequences for modelling and understanding the phenomenon at hand. It also may reduce the generality and versatility of the macro-model to examine other closely related phenomena. Van Assen (2001: 175–178) explains this in the context of exchange.

Individuals and their characteristics play no role in split pool exchange, as they are assumed away too. But where then, does the exchange relation come from? The network structure is *exogenously* determined. In pure exchange, as in the experiments of Michener et al. (1975, 1977) potential exchange relations and the exchange network are *endogenously* determined by individuals' endowments and preferences over these endowments. Consequently, effects of networks on exchange outcomes may be naturally examined using the more complex model of pure exchange rather than split pool exchange. Similarly, the *evolution of exchange (networks)* may naturally be studied by providing a group of actors with endowments and preferences and allow them to interact with each other repeatedly, without restrictions such as in the Michener et al. experiments. Changes in endowments or preferences may induce a change in the actual network of exchanges.

Characterizing actors by endowments (or characteristics) and preferences also allows for naturally modelling *generalized exchange*, or other social interactions such

as *coercion* and *conflict*. In generalized exchange the exchange occurs between three of more actors and cannot be reduced to exchange between two actors. For instance, A benefits form a transfer of goods by B, B profits from a transfer by C, and C profits from a transfer by A. Some profitable exchanges in Michener et al.'s (1975, 1977) experiments could only be realized using generalized exchange (see Dogan and Van Assen, 2009), revealing that empirical research on generalized exchange can be naturally conducted using pure exchange. Generalized exchange cannot be modelled using split pool exchange. Similarly, it is not obvious how coercion and conflict can be researched using split pool exchange. In coercive relations one actor has the potential to unilaterally transfer a resource that harms the other, whereas in conflict relations both actors have this potential. Both coercive and conflict relations are studied in sociology with pure exchange (Willer 1999).

Another example of an application of pure exchange that is not obvious using split pool exchange is *exchange with externalities*. Externalities of exchange occur when an exchange between two actors also affects the utility of one or more actors that or not involved in the exchange. Life is abundant with examples of exchanges with externalities, for instance when an actor exchanges on behalf of a group to which (s)he belongs (for instance, household, company, political party). Using pure exchange, Dijkstra and Van Assen (2006, 2008b) demonstrate that externalities greatly affect outcomes and behavior of actors embedded in a simple exchange network. Additionally, pure exchange with externalities is effectively used to predict bargaining and outcomes of collective decision making (Dijkstra, Van Assen, and Stokman 2008; Van Assen, Stokman, and Van Oosten 2003).

Pure exchange with externalities can also be naturally applied to examine social dilemmas. The Prisoner's Dilemma game (PD) is one of the most frequently used paradigms for studying cooperation between individuals. Dijkstra and Van Assen (2016) argue that the PD is often an inadequate and too simple model of social dilemmas, violating the principle of sufficient complexity, just like split pool exchange is a too simple model of exchange. That is, the PD assumes away third parties that generate the dilemma (for instance, consumers of fish who may offer more money to buy fish, thereby tempting the fishermen to defect, that is, to sell fish), the fact that outcomes of a social dilemma are usually variable and negotiable, and that behavior and outcomes are sequential rather than simultaneous (see Dijkstra and Van Assen, 2016, for details). Modelling the social dilemma as an *exchange dilemma* (pure exchange with externalities) instead of a PD naturally incorporates these aspects of a social dilemma. Interestingly, Dijkstra and Van Assen (2016) demonstrate empirically that outcomes of a social dilemma modelled as an exchange dilemma may result in very different outcomes than typically obtained with a PD; cooperation in the exchange dilemma was very rare, which is in contrast with cooperation rates typically observed in research using the PD. Their results again induce the fundamental question if outcomes of research using the PD are ecologically valid for real-life social

dilemmas, or that the PD is a too simple model of the social dilemma assuming away essential ingredients of the social dilemma.

We hope that our paper increases sociologists' sensitivity to the principle of sufficient complexity when modeling social phenomena. Using the case of exchange we showed that too simple a model (split pool exchange) assumes away essential ingredients of exchange, making it impossible studying real-life exchange and these ingredients. Unsurprisingly, assuming away these ingredients greatly affects both predictions and outcomes of exchange, unavoidably inducing the question if outcomes of and conclusions based on research using the too simple model are valid for the phenomenon at hand. Applying the principle of sufficient complexity may imply that the complexity of the model (micro, macro, or both) depends on the social phenomenon one tries to understand. For instance, although we argue that in many cases of exchange the pure exchange model is superior to the more commonly used split pool exchange, there may be cases in which the split-pool exchange model is preferred or an even more complex model than pure exchange is warranted.

References

Bonacich, Phillip. 1992. "Power in Positively Connected Exchange Networks: A Critical Review." *Advances in Group Processes 9*: 21–40.

Burt, Ronald S. 2005. *Brokerage and Closure: An Introduction to Social Capital*. Oxford: Oxford University Press.

Camerer, Colin F. 2003. *Behavioral Game Theory: Experiments in Strategic Interaction*. Princeton, NJ: Princeton University Press.

Coleman, James S. 1990. *Foundations of Social Theory*. Cambridge, MA: Harvard University Press.

Coleman, James S. 1987. "Microfoundations and Macrosocial Behavior." Pp.153–173 in *The Micro-Macro Link* eds. Jeffrey C. Alexander, et al. Berkeley: University of California Press.

Cook, Karen S. and Emerson, Richard M. 1978. "Power, Equity and Commitment in Exchange Networks." *American Sociological Review* 43: 721–739.

Dijkstra, Jacob and Van Assen, Marcel A. L. M. 2006. "Effects of Externalities on Exchange in Networks." *Sociological Theory and Methods* 21(2): 279–294.

Dijkstra, Jacob and Van Assen, Marcel A. L. M. 2008a. "Transferring Goods or Splitting a Resource Pool." *Social Psychology Quarterly* 71(1): 17–36.

Dijkstra, Jacob and Van Assen, Marcel A. L. M. 2008b. "The Comparison of Four Types of Everyday Interdependencies: Externalities in Exchange Networks." *Rationality and Society* 20(1): 115–143.

Dijkstra, Jacob and Van Assen, Marcel A. L. M. 2016. "Social Dilemmas as Exchange Dilemmas." *Corvinus Journal of Sociology and Social Policy* 7(2): 55–76.

Dijkstra, Jacob, Van Assen, Marcel. A. L. M. and Stokman, Frans N. 2008. "Outcomes of Collective Decisions with Externalities Predicted." *Journal of Theoretical Politics* 20(4): 415–441.

Dogan, Gonul and Van Assen, Marcel A. L. M. 2009. "Testing Models of Pure Exchange." *Journal of Mathematical Sociology* 33(2): 97–128.

Edgeworth, Francis Y. 1881. *Mathematical Physics*. Fairfield: Kelley.

Friedman, Milton. 1953. *Methodology of Positive Economics*. Chicago: University of Chicago Press.
Hedström, Paul 2005. *Dissecting the Social: On the Principles of Analytical Sociology*. Cambridge University Press, Cambridge.
Hildenbrand, Werner and Kirman, Alan P. (1988). *Equilibrium Analysis*. Amsterdam, Netherlands: North-Holland.
Homans, George C. 1958. "Social Behavior as Exchange." *Journal of American Sociology* 62: 597–606.
Kalai, Ehud and Smorodinsky, Meir 1975. "Other Solutions to Nash's Bargaining Problem." *Econometrica* 43: 513–518.
Kiss principle. (n.d.). In Wikipedia. Retrieved December 11, 2019, from https://en.wikipedia.org/wiki/KISS_principle.
Lindenberg, Siegwart. 2001. "Social Rationality versus Rational Egoism." Pp. 635–668 in *Handbook of Sociological Theory*, ed Jonathan H. Turner. New York: Kluwer Academic/Plenum Publishers.
Michener, H. Andrew, Cohen, Eugene D. and Sørensen, Aage B. 1975. "Social Exchange: Predicting Interpersonal Outcomes in Four-Event, Three Person Systems." *Journal of Personality and Social Psychology* 32: 283–293.
Michener, H. Andrew, Cohen, Eugene D. and Sørensen, Aage B. 1977. "Social Exchange: Predicting Transactional Outcomes in Five-Event, Four Person Systems." *American Sociological Review* 42: 522–535.
Nash, John F. 1950. "The Bargaining Problem." *Econometrica* 18: 155–162.
Roth, Alvin E. 1995. "Introduction to Experimental Economics" Pp. 3–109 in *The Handbook of Experimental Economics* eds John H. Kagel and Alvin E. Roth. Princeton, N.J.: Princeton University Press.
Shubik, Martin 1982. *Game Theory in the Social Sciences: Concepts and Solutions*. Cambridge, MA: MIT Press.
Van Assen, Marcel. A. L. M. 2001. *Essays on actor models in exchange networks and social dilemmas*. (Doctoral dissertation, University of Groningen).
Van Assen, Marcel. A. L. M. 2003. "Exchange Networks: An Analysis of All Networks up to Size 9." Pp. 67–103 in *Power and Status*. Oxford: Emerald Group Publishing Limited.
Van Assen, Marcel A. L. M., Stokman, Frans N. and Van Oosten, Reinier 2003. "Conflict Measures in Cooperative Exchange Models of Collective Decision-Making." *Rationality and Society* 15(1): 85–112.
Willer, David ed. 1999. *Network Exchange Theory*. Westport, CT: Praeger Press.

Andreas Flache

5 Rational Exploitation of the Core by the Periphery? On the Collective (In)efficiency of Endogenous Enforcement of Universal Conditional Cooperation in a Core-Periphery Network

Abstract: Raub and Weesie (1990) proposed a game theoretical model addressing effects of network embeddedness on conditional cooperation between two actors. This work showed that network embeddedness can facilitate conditional cooperation by reducing uncertainty, in line with a number of follow-up contributions and consistently with results from other modelling approaches. This research focused mainly on interactions between two parties embedded in a network. In the present paper, I extend a closely related model towards *N*-person collective good problems, combining conditional cooperation based on direct monitoring via network ties and observation of group output in an uncertain environment. The focus is on a maximally simple yet empirically relevant case, a core-periphery network in which only core-members can directly observe each other's contributions to a collective effort, whereas peripheral members only observe a noisy signal indicating aggregated contributions. I propose the possibility of a 'rational exploitation of the core by the periphery'. Strategy-profiles in which free-riding of peripheral members is tolerated while core-members cooperate conditionally, are not only individually rational but also payoff-superior to profiles with universal conditional cooperation if uncertainty is sufficiently high and the number of peripheral members is sufficiently low.

5.1 Introduction

The question under which conditions cooperation can be achieved in the production of collective goods is still prominently on the agenda of social scientists more than five decades after Olson's (1965) influential analysis of the problem of collective action. Especially for "large number dilemmas" (Raub 1988), Olson and many

Notes: I am greatly indebted to Werner Raub for the inspiration and guidance over the years that has formed the basis for this work and for so much more. I also thank the editors for a thorough review and constructive comments on an earlier version of the manuscript.

Andreas Flache, University of Groningen, Department of Sociology / ICS

https://doi.org/10.1515/9783110647495-005

other authors (for example Hardin 1968) were pessimistic about the possibility of an endogenous solution without hierarchical enforcement. However, game theoretical research has shown that reciprocity in the form of conditional cooperation can be an individually rational endogenous solution of collective action problems (Friedman 1971, 1986; Raub and Voss 1986; Raub 1988; Taylor 1976, 1987) in a collective action situation that constitutes an infinitely or indefinitely repeated game.

Conditional cooperation in collective good production makes an actor's contribution to the common effort contingent upon others' contributions in the past, similar to the well-known strategy of "Tit-for-Tat" (Axelrod 1984) in 2-person social dilemma games. An important problem that can limit the effectiveness of conditional cooperation is that actors may not be able to observe precisely and reliably others' past contribution (Bendor and Mookherjee 1987; Green and Porter 1984). For example, when a project team member in an organization fails to show up for a project meeting, this may be intended free-riding and should thus trigger retaliation from other team members to credibly deter future free-riding according to the logic of conditional cooperation. But possibly the absent team member was delayed by an incident he cannot be held accountable for, like unexpected sickness of a child. What makes things worse, it is often hard to verify for other actors involved what the true reasons were if someone failed to contribute, or who in a team was responsible when the team's results failed to meet expectations. The dilemma here is that sanctions are needed in such a situation to deter free-riding behavior. But such sanctions also bear the danger of putting successful cooperation under pressure, for example when the target of a sanction feels treated unfairly and responds with a counter-sanction (Nikiforakis 2008; Nikiforakis and Engelmann 2011) potentially evoking cycles of mutual recrimination (Bendor, Kramer and Stout 1991; Kollock 1993; Wu and Axelrod 1995) disrupting ongoing cooperation more than necessary.

Network embeddedness (Granovetter 1985) can mitigate the difficulties that arise from information problems in ongoing collective action. Network ties connecting some of the participants of the collective action more closely than others, such as friendship relations between some team members, or close physical proximity of workplaces or homes, give some the opportunity to observe more reliably whether their network contacts contributed to a joint effort or for which reasons they failed to do so. Moreover, network ties allow to communicate this information to further participants of the collective action. As a result, a network of interpersonal ties may stabilize conditional cooperation in a situation where contributions are hard to observe without direct network connections, because either through direct observation or communication via network relations participants know better why failures to contribute occurred and sanctions can thus be more effectively directed at free riders.

Raub and Weesie (1990) developed a game-theoretical model formalizing this argument and showed how and under which conditions social networks facilitate trust in cooperation problems faced in two-party relations (see also Raub and Weesie,

2000), as they occur for example in business relations or between partners in a household. A range of follow-up studies extended this analysis and tested the argument empirically (Batenburg, Raub and Snijders 2003; Buskens 2002; Buskens and Raub 2002; Raub 2017; Raub and Buskens 2008; Rooks, Raub, and Tazelaar 2006). This work contributed to a wider literature of models that link network ties to cooperation. Most of that literature focuses on mechanisms other than conditional cooperation. Some examples are imitation (Gould 1993), threshold dynamics (Chwe 1999; Macy 1991), mobilization via network ties (Marwell and Oliver 1993), sanctions imposed via social ties (Coleman 1990; Flache 1996; Flache, Macy, and Raub 2000), or information on others' preferences acquired via network relations (Dijkstra and van Assen 2013). Some authors also modelled conditional cooperation in repeated collective good games (Bednar 2006; Bendor and Mookherjee 1990; Flache 2002; Flache et al. 2000; Spagnolo 1999; Wolitzky 2013). Especially Fatas and co-authors (for example Fatas, Meléndez-Jiménez, and Solaz 2010; Fatas et al. 2015) developed a line of papers combining theoretical modelling and experimental tests of effects of networks on conditional cooperation in collective good settings.

Despite some exceptions, the problem of cooperation in collective goods has received relatively little attention in the research that links network embeddedness to the feasibility of conditional cooperation under uncertainty, compared to the problem of cooperation in two-party relations. One possible reason is that models of conditional cooperation in collective action typically assume that in addition to networks there is another source of information actors can rely upon to condition their own contribution behavior on others' cooperation. This is the level of the good provided, indicating how many group members made a contribution to bring it about. A number of theoretical studies (Friedman 1971; Raub 1988; Raub and Voss 1986; Taylor 1987) highlighted that this information is under certain conditions sufficient to render it an individually rational endogenous outcome of the game if everyone cooperates conditionally upon sufficient group-output. But imperfect information provides an important problem also for this endogenous solution. Observed levels of provision of a collective good are rarely a perfect or reliable indicator of how much effort group members really have invested to bring the good about. Bendor and Mookherjee (1987; see also Bendor, Kramer and Stout 1991; Kollock 1993) prominently showed how this can make conditional cooperation highly inefficient. If the condition for cooperation is sufficient group-output, this implies that under uncertainty sometimes the condition may not be met for reasons which are unrelated to deliberate free-riding by some group members – like the unintended failure of a team member to turn up for a project meeting. Rational actors face the dilemma that therefore sanctioning strategies must be lenient to some extent to avoid too much 'unnecessary' mutual punishment, but that too much lenience invites deliberate defection because deviants can hope to 'get away' with occasional free-riding. In other words, the social costs of enforcing cooperation by sufficiently severe sanction threats may become prohibitively high if there is too much uncertainty about

the link between efforts actors make to contribute to a collective good and the observable results in terms of the level of provision.

Work on the effect of uncertainty for conditional cooperation suggests that monitoring in social networks can be an important factor that stabilizes conditional cooperation in collective action. Yet, hitherto only few studies address the role of monitoring via network ties in a setting where actors simultaneously observe a group-output that is an unreliably indicator of actual contributions. In this paper, I try to take a step in that direction by proposing a game-theoretical model of collective action under uncertainty, aiming to integrate effects of monitoring via social network ties albeit with a maximally simple network structure.

In modelling the network structure in a maximally simple way, I want to explore a structural problem that occurs in many situations where collective action is needed. Often, networks are heterogeneous in the extent to which their members are connected among each other.

One particular case of heterogeneity are core-periphery structures, in which some members of an interest group form a densely connected core and others are in peripheral positions with only sparse links to members in the core. Consider as an extremely simple case the situation in which core members can monitor each other's contributions to the collective effort very closely and accurately, while the effects of the contributions of peripheral members are only visible via their impact on the group-output, while their real efforts to contribute to the collective good are private information. Empirical settings for which this could be seen as an ideal-typical model of a collective good situation might be (a) a company that works with staff located in headquarters and local representatives dispersed across different regions or countries, (b) a semi-virtual organization, working with a local core team of members physically located in an office, and a number of workers who are only connected online with each other and the core-members (Flache 2004), or (c) a local renewable energy initiative largely driven by a core-team of densely connected 'front-runners' but in need of contributions from a larger number of members of the community who are much less connected with the front-runners than they are connected with each other (Goedkoop, Flache, and Dijkstra 2017).

Core-periphery structures can impose a dilemma for endogenous conditional cooperation. Under uncertainty, sanctioning regimes that provide sufficient incentives for all members to cooperate in equilibrium cannot avoid that some punishment must be imposed even when only 'erroneous' defections occur, because with imperfectly observed input of the peripheral members their free riding can otherwise not be credibly deterred. In this situation the closer embeddedness of the core-members can make an alternative regime more attractive, a sanctioning regime in which core-members ignore the unreliable and noisy information about collective output, but condition their behavior instead on the reliably observed inputs of only the other core-members. Obviously, the problem with that alternative solution is that peripheral members can no longer be credibly deterred from free riding. They are effectively

allowed to take a free-ride, an outcome that I call here exploitation of the core by the periphery, mirroring Olson's (1965) 'exploitation of the big by the small'. What makes this possibility nevertheless worth to investigate is that it may under certain conditions be collectively more efficient than a conditional cooperation regime in which sanctions can be triggered by everyone's failure to contribute. Intuitively, the reason is that credible enforcement of the cooperation of those group members whose actions are difficult to observe is not possible without risking a considerable amount of sanctions that are triggered by unintended failures to generate effective contributions. Imposing such a sanction is damaging and potentially disruptive in itself for the collective effort. Tolerating some free riding from peripheral members avoids these costs, potentially rendering such tolerance a collectively more desirable solution. As an example, consider the renewable energy initiative discussed above. If every time when not enough community members show up for an information meeting about the initiative, the core-team of front-runners would stop its activities for some time to sanction 'free-riders' in the community, the initiative may suffer more damage from those sanctions than was caused by the lack of contribution of peripheral community members in the first place. Accepting relatively low turn-out at information meetings may thus be a collectively more efficient strategy of the front-runners than trying to enforce contributions from all community members all the time.

In what follows I investigate the conditions under which the outcome in which peripheral members free ride and core-members contribute can be both individually rational as well as socially more efficient than universal conditional cooperation. The model I use for this will be described in Section 5.2, results for specific scenarios are presented in Section 5.3 and the paper closes with a discussion of possible implications and limitations in Section 5.4.

5.2 Model

In section 5.2.1. the repeated game is presented, section 5.2.2. describes the approach for analysis of the conditions for individual rationality and social efficiency of the alternative outcomes of universal conditional cooperation on the one hand, and "core-only" conditional cooperation on the other hand.

5.2.1 The repeated game

Group interaction is modelled as a repeated N-person game that is equivalent to the "work game" in Flache (2002), except for the assumption that monitoring of contribution behavior is possible via direct network-ties. The constituent-game strategy of player i in iteration t of the repeated game is represented by the decision w_{it}

whether to "work" or "shirk", where $w_{it} = 0$ for defectors and $w_{it} = 1$ for contributors. In the stage game, actors take decisions simultaneously and independently.

Following Bendor and Mookherjee (1987) uncertainty is modelled with a commonly known universal probability ε that due to some mishap an individual's contribution fails to be effective ($0 \leq \varepsilon \leq 1$). It is assumed that all players know after every iteration t the *group-output* in terms of the number of *effective* contributions group members made in t, but players are not necessarily aware of the actual input w_{jt} of other individual members.

Modelling the core-periphery structure, I assume a maximally simple social network S. The group consists of N_c core-members and N_i isolates ($N = N_c + N_i$). Core-members are connected to all other core-members in the network, whereas isolates are not connected to anyone else. If there is a connection between individuals i and j ($s_{ij} = 1$), i is at time t, before taking her own decision, fully and perfectly informed on the actual contribution decision w_{jt} her network contact j made in all previous iterations $t' < t$. This is a rather extreme simplification, but it greatly facilitates model analysis, while it still captures the substantive assumption that monitoring contributions via network relations is more reliable than between unconnected actors.

The expected payoff of actor i in iteration t of the game, u_{it}, results from both expected benefits from output and expected costs of i's own contribution-effort. Output is simply modelled as a linear function of the sum of individual outputs. Notice that output may be lower than the number of actual contributions made, due to uncertainty. The expected output generated by an individual contribution is $(1 - \varepsilon)\, w_{it}$. The amount of the collective good produced is shared equally among all group members, whereas costs of making a contribution are private. Equation (5.1) formalizes the expected payoff an individual derives from the outcome of the constituent game in iteration t:

$$u_{it} = \sum_{j=1}^{N} \left(\frac{\alpha}{N} (1 - \varepsilon) w_{jt} - c w_{it} \right). \tag{5.1}$$

The parameter α scales the benefit a group member receives from consuming a unit of the collective good. The costs of investing a unit of effort into its production are indicated by the parameter c. Modelling a problematic collective action situation, the constituent game has a N-person Prisoner's dilemma structure given by $\alpha / N < c < \alpha$.

To model conditional cooperation in the repeated game, I use the standard assumption of infinite repetition of the game with exponential discounting of future payoffs. The accumulated payoff u_i of actor i in the repeated game sums discounted payoffs over all iterations t, u_{it}. Formally,

$$u_i = \sum_{t=0}^{\infty} \tau^t u_{it}, \quad 0 < \tau < 1. \tag{5.2}$$

where τ is the discount parameter, defining the value of players' interest in future payoffs. For simplicity, τ as well all the parameters α, ε and c are assumed equal for all members of the group.

5.2.2 Model analysis

5.2.2.1 Strategy types

Main aim of the model analysis is comparison of the conditions for individual rationality and social efficiency of two types of strategy profiles for the repeated game, modelling two different types of reciprocity norms. In this game, the only 'weapon' group members have to impose a sanction on defectors is response with own defection. Conditional cooperation thus implies reciprocal behavior in the sense that all players cooperate as long as there has been sufficient cooperation by others in the past, but resort to defection otherwise. The first type of strategy profile analyzed here meets the additional requirement of symmetry. The stage-game behavior expected from players, the attached conditions for triggering a sanction and the severity of the sanction are universally shared in this profile. I call the corresponding strategy-type "universal conditional cooperation" in what follows. Especially, both core-members and isolates follow the same reciprocity norm under universal conditional cooperation. This can be seen as a property that ensures 'procedural fairness', a key ingredient for sustainable cooperation according to many authors. The second type of strategy profile relaxes the 'procedural fairness' requirement of symmetry for the sake of potential gains in social efficiency. This type differentiates between the norm imposed upon core-members and a maximally lenient norm imposed upon isolates, effectively allowing them to free-ride. I call this type "core-only" conditional cooperation. Both strategy-types as well as their analysis are discussed in more detail below.

5.2.2.2 Individual rationality and social efficiency of universal conditional cooperation

A key principle for the selection of strategy profiles as solutions of the game is *individual rationality*. Technically, this implies that the solution of the game is a subgame perfect Nash equilibrium (s.p.e.) (Selten 1965; see also Kreps 1990). Another key principle is *social efficiency* in terms of payoff dominance. Payoff dominance eliminates those s.p.e.'s from the set of possible solutions of a game to which all players would unanimously prefer other s.p.e.'s (for more details, see Harsanyi 1977:116–119). Among the set of *individually rational and symmetric reciprocity strategy profiles*, those will be selected as candidate-solution of the game that yield a higher payoff to all members of the group compared to other strategy-profiles in

this set. Given symmetry, this means the s.p.e. will be selected that maximizes u_i as defined by (5.2) for all players.

The analysis uses a generalized form of trigger strategies (Friedman 1971, 1986), following Bendor and Mookherjee (1987; see also Flache 2002). Trigger strategies under imperfect information generate cooperative behavior in a *normal period* of the game, but as soon as the corresponding group-output norm has been violated, the trigger strategy reverts to a punishment behavior for a subsequent *sanctioning period*. After the sanctioning period, cooperation is restored but only as long as there is sufficient group-output. Two parameters of a trigger strategy model the tradeoff between lenience and deterrence, the cutoff level l and the sanction time s. In the strategy-profile $\sigma(s,l)$ all players cooperate in all rounds in a normal period, but they revert to a sanctioning period of exactly s rounds, as soon as in the normal period the group-output falls below the cut-off level l. After the sanctioning period, a new normal period starts with an initial round of unconditional cooperation.

To guarantee individual rationality of adhering to $\sigma(s, l)$, the profile needs to ensure that optimal unilateral deviations from it do not pay. To also optimize efficiency, the strategy profile $\sigma(s^*, l^*)$ is sought that maximizes the related expected payoff $u_i(\sigma(s^*, l^*))$, subject to the constraint that the corresponding trigger strategy constitutes a subgame perfect equilibrium.[1]

For calculation of expected payoffs of trigger-strategies and for evaluation of the constraint that unilateral deviations do not pay, I adapted an efficient *numerical* algorithm from Bendor and Mookherjee (1987) that solves the optimization problem for a given set of conditions. (α, c, N, N_c, ε, τ). The algorithm will here only be sketched in broad strokes, more details can be found in Flache (2002). The analysis of the trigger strategy profile "task-cooperation" in Flache (2002:199) is equivalent to the analysis of $\sigma(s,l)$ here.

The algorithm finds the payoff-dominant s.p.e. in the set of $\sigma(s,l)$ profiles in two steps. In the first step, it is established for which cut-off levels l' individually rational profiles $\sigma(s,l')$ can exist. Such a profile exists if and only if the condition is satisfied that the trigger strategy $\sigma(\infty,l')$ with *eternal punishment* and cut-off level l' is

1 Due to symmetry, trigger strategies always constitute a subgame perfect equilibrium, if and only if they satisfy the conditions for Nash-equilibrium. The proof is given by Friedman (1986). For the game at hand here a possible complication is that core-members have more information about possible deviations than isolates have. Off the equilibrium path, a core-member can observe deviations by other core-members even if those deviations do not lead to a sufficient drop in output to trigger a subsequent punishment phase under $\sigma(s^*, l^*)$. The core-member might thus have an incentive to deviate from the profile $\sigma(s^*, l^*)$ in the subgame that ensues, because the observed deviation changes the probability that there will be a punishment phase in the next round. To avoid such complications I assume that both group-output and individual inputs become known only after all stage-game decisions were taken and before the next round begins. This makes it public knowledge for all group-members whether in the subsequent round a punishment phase will start, aligning the conditions for deviation in the subgame with those for deviation in the overall game.

individually rational. Intuitively, the reason is that eternal punishment maximizes the expected loss from the sanction that a deviant faces. If eternal damnation is not sufficient to deter deviation at l', no finite sanction time s is. Theorem 1 in Flache (2002) gives the condition that follows, which in turn is efficiently tested with the numerical algorithm for all feasible cut-off levels l', $1 \leq l' \leq N$.

The second step of the numerical procedure is to calculate and compare the optimal expected payoffs from universal conditional cooperation that can be obtained for those cut-off levels l' that satisfy the condition that $\sigma(\infty,l')$ is s.p.e. The optimal profiles $\sigma(s',l')$ can be found by inspection of only one trigger strategy per cut-off level l', the one that minimizes sanction time s, subject to the constraint of individual rationality $u_i - u_{-i}$ under $l = l'$,where u_i and u_{-i} denote the payoffs for universal cooperation and the optimal unilateral deviation from that profile for i, respectively. Shorter sanctioning periods ensure higher payoffs for all under universal conditional cooperation. But there is also a critical lowest sanction time s^*, below which the individual rationality constraint $u_i \geq u_{-i}$ can no longer be satisfied because sanctions become too lenient. Theorem 2 in Flache (2002) specifies how this sanction-time s^* is efficiently computed by the numerical algorithm.

Given a vector of parameters (α, c, N, N_c, ε, τ) of the game, the algorithm finds the payoff-dominant s.p.e. $\sigma(s^*(l'),l')$ for every cut-off level l' for which an individually rational profile $\sigma(s,l')$ exists. Among these payoff-dominant profiles $\sigma(s^*(l'),l')$, the algorithm then selects the cut-off level l^* that maximizes the related expected payoff. If no other s.p.e. exists for the given set of parameters, the unique symmetric solution of the game is universal and full defection.

5.2.2.3 Individual rationality of core-only conditional cooperation

The core-only strategy-profile σ_{co} implies unconditional contribution to the collective good in the first round of the game for core-members. Thereafter, core-members contribute if and only if all other core-members have contributed in all previous rounds of the game. That is, as soon as a core-member detects defection by another core-member, all core-members revert to eternal defection. Core-members condition their behavior exclusively upon the observed actions of other core-members. In what follows, the analysis will be restricted to the most extreme form of a "core-only" strategy-profile, the form in which all isolates choose full defection. The conditions under which such a profile is individually rational for core-members can be seen as the most restrictive conditions under which individually rational profiles exist at all that demand less cooperation from isolates than from core-members. If core-members are willing to adopt this core-only strategy, strategy profiles with higher levels of cooperation by isolates would yield higher payoffs for all core members and thus also be individually rational for them. Alternatively, isolates could adopt a $\sigma(s,l)$ profile in which they would enforce some level of conditional cooperation from each other with the

threat of sanctioning if group-output drops too low. Yet, at least as long as isolates are in the minority, this would impose only a relatively weak sanction because group-output drops only by $(1 - \varepsilon)N_i$. in sanctioning periods. This suggests that conditions under which such a profile is individually rational also for isolates are rather restrictive and demand very long sanctioning periods, which diminishes possible efficiency gains compared to the extreme case of full defection by isolates.

The condition for individual rationality of cooperation among only the core-members in the strategy-profile σ_{co} follows from Friedman's theorem (1971; 1986: 88–89) for indefinitely repeated games with perfect monitoring of past behavior. σ_{co} constitutes a subgame perfect equilibrium, if and only if core-members are sufficiently interested in future payoffs. This the condition given by equation (5.3).

$$\tau > \tau^* = \frac{\hat{T}_{co} - \hat{R}_{co}}{\hat{T}_{co} - P}. \tag{5.3}$$

The symbol $\hat{T}_{co}$ denotes the expected stage-game payoff from unilateral deviation by a core-member, $\hat{T}_{co} = (1-\varepsilon)\alpha(N_c - 1)/N$. The symbol $\hat{R}_{co}$ refers to the stage-game payoff for a core-member of universal cooperation by only core-members, $\hat{R}_{co} = (1-\varepsilon)\alpha N_c/N - c$. P, finally, is the stage-game payoff of universal defection for a core-member, which is zero.

5.2.2.4 Solution of the game

The solution of the game that is selected for a given parameter-vector (α, c, N, N_c, ε, τ) is full defection by all players if neither any $\sigma(s, l)$ nor σ_{co} constitute an s.p.e. If only σ_{co} constitutes an s.p.e. this is the solution, if only $\sigma(s,l)$ profiles are s.p.e. then the payoff-dominant solution $\sigma(s^*, l^*)$ is the solution. Finally, if both $\sigma(s,l)$ profiles and σ_{co} constitute s.p.e.'s, then the payoff dominant one among the set of all these s.p.e.'s is selected as solution. Notice that for σ_{co} this requires in particular that the accumulated payoff is higher both for core-members and for isolates compared to the payoff all members receive in $\sigma(s^*, l^*)$. Accumulated payoff as well as the incentive to deviate, $u_i - u_{-i}$ for σ_{co} can be obtained from $u_{it} = \hat{R}_{co}$, $u_{-it} = \hat{T}_{co}$ and equation (5.2). Accumulated payoff u_i and incentive to deviate, $u_i - u_{-i}$ for $\sigma(s,l)$ profiles are computed as given in Flache (2002).

5.3 Results

First, it will be explored for an illustrative scenario how changes in the 'shadow of the future' τ, the number of isolates N_i, and the degree of uncertainty ε affect the conditions for individual rationality and efficiency of a specific trigger-profile $\sigma(s^*,l^*)$

compared to σ_{co}. Second, comparative statics will be reported about the effects of these variables on the distribution of optimal solutions across the entire range of possible values of τ.

5.3.1 Illustrative scenario

The scenario inspected here represents a relatively small work-team with a minority of isolates, facing a moderately severe cooperation problem under considerable uncertainty, formalized as $N = 10$; $\varepsilon = 0.2$; $\alpha = 1$; $c = 0.33$; $N_i = 2$. I choose the trigger-profile $\sigma(s^*, l^*)$ that under these conditions is payoff-dominant among all $\sigma(s, l)$ profiles at $\tau = 0.95$. This is $s^* = 6$ and $l^* = 33$.

5.3.1.1 Effects of shadow of the future (τ)

The feasibility of conditional cooperation depends crucially upon sufficient interest of actors in future outcomes, τ. In the following, it will be analyzed how changes in τ affect the individual rationality and the payoffs core-members obtain for the profile $\sigma(s^*,l^*)$ and for σ_{co}. Figure 5.1a shows how changes in τ affect the payoff-difference $u_i - u_{-i}$ for core-members between universal adherence to the trigger strategy and the optimal unilateral deviation for both $\sigma(s^*, l^*)$ and σ_{co}. Only when this difference is above the zero line, the corresponding profile is an s.p.e. Figure 5.1b shows corresponding changes in the accumulated payoff u_i of core-members in the two strategy-profiles.

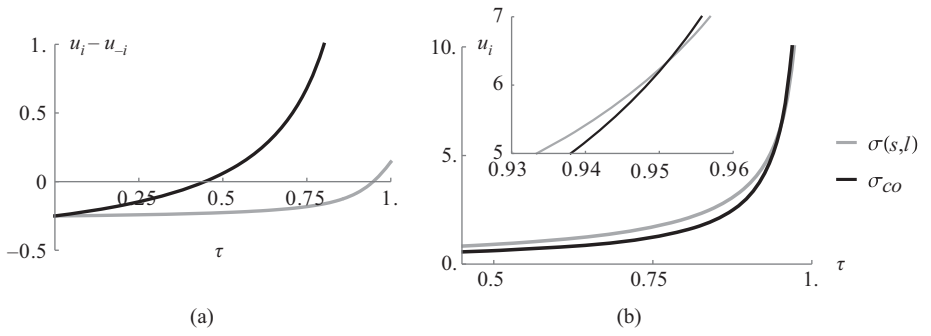


Figure 5.1: (a): Effect of τ on payoff difference universal trigger vs. unilateral deviation $u_i - u_{-i}$. **(b)**: Effect of τ on payoff u_i for core-members. Inset: range where payoff curves intersect.

For a significant range of τ (between approximately $\tau = 0.45$ and $\tau = 0.95$) core-only cooperation is found to be payoff inferior to universal cooperation under $\sigma(s^*,l^*)$, yet universal cooperation is not individually rational in this range while core-only

cooperation is. Only when τ exceeds a critical threshold of about $\tau = 0.95$ both profiles are individually rational and universal conditional cooperation becomes the payoff superior solution. However, as τ further increases, a point is reached where core-only cooperation becomes payoff-superior to universal cooperation (see inset Figure 5.1b). This reflects that if interest in the future is sufficiently strong, the loss of output caused by free riding of two isolates is outweighed by the gain from avoiding frequent punishment periods.

5.3.1.2 Effects of number of isolates (N_i) and of noise (ε) in baseline-scenario

The baseline scenario suggests that core-only cooperation is the only sustainable solution for a considerable range of conditions and even becomes the most efficient solution when τ is high enough. The main advantage of the core-only solution is the avoidance of frequent punishment phases caused by uncertainty, while its main disadvantage is the loss of the contributions of peripheral members. Intuitively, a larger share of isolates should then reduce the relative attractiveness of a core-only solution, while more uncertainty should increase it. In the following it will be demonstrated that the game-theoretical model is consistent with these intuitions for the baseline scenario. Figure 5.2 shows how the number of isolates N_i affects the individual rationality and efficiency of $\sigma(s^*,l^*)$ and of σ_{co} at the level of $\tau = 0.95$ and how this effect interacts with a change in the degree of uncertainty from $\varepsilon = 0.2$ to $\varepsilon = 0.25$. As $\sigma(33, 6)$ is s.p.e. throughout in Figure 5.2, the figures show only the accumulated payoff u_i for this profile. For σ_{co} both the payoff-difference $u_i - u_{-i}$ and the corresponding accumulated payoff u_i for core-members are shown. Only when $u_i - u_{-i}$ is above zero, σ_{co} is an s.p.e., only when u_i for σ_{co} is above u_i for $\sigma(s^*, l^*)$ in figure 5.2, core-only conditional cooperation is also payoff-superior to universal conditional cooperation.

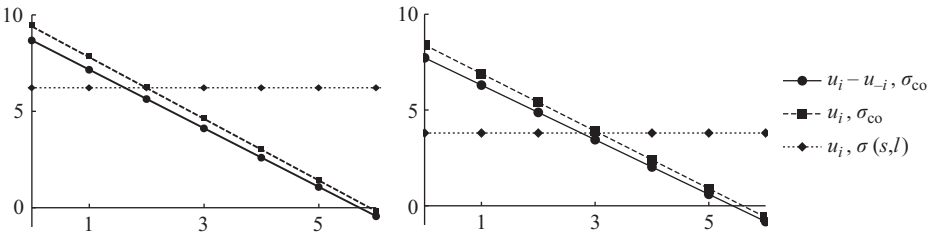


Figure 5.2: Effect of N_i on accumulated payoff for both $\sigma(s^*, l^*)$ and σ_{co}, and effect on $u_i - u_{-i}$ for σ_{co}. Baseline scenario $N = 10$; $\alpha = 1$; $c = 0.33$; $\tau = 0.95$, $s = 33$, $l = 6$. (a) $\varepsilon = 0.2$, (b): $\varepsilon = 0.25$.

Figure 5.2 demonstrates how core-only conditional cooperation can be both individually rational and payoff-superior to universal conditional cooperation, but only as long as the number of isolates is sufficiently small. For $\varepsilon = 0.2$, σ_{co} is individually

rational for $N_i \leq 5$, and it is payoff-superior to $\sigma(s^*, l^*)$ for $N_i \leq 1$. If the level of uncertainty increases to $\varepsilon = 0.25$, core-members benefit less from cooperation in σ_{co}, but also efficiency-losses of $\sigma(s^*, l^*)$ become higher. As a consequence, the range within which σ_{co} is the solution of the game widens considerably. The range within which σ_{co} is both individually rational and payoff-superior increases from $N_i \leq 1$ for $\varepsilon = 0.2$ to $N_i \leq 3$ for $\varepsilon = 0.25$. Notice that based on condition (5.3) it is straightforward to prove analytically for all feasible parameter vectors that s.p.e. conditions for σ_{co} become more restrictive both for increasing N_i as well as for increasing ε. I cannot offer a corresponding analytical result for the payoff-superiority of σ_{co} over $\sigma(s^*,l^*)$ but numerical analyses suggest that the results shown here for the baseline scenario generalize to a much wider range of conditions.

5.3.2 Comparative statics

5.3.2.1 Overview

In this section comparative statics will be reported about the effects of uncertainty, ε, and the number of isolates, N_i, on the range of possible values of τ for which either the optimal $\sigma(s,l)$ profile $\sigma(s^*, l^*)$, or σ_{co} or full defection is the solution of the game. The parameters that are fixed are again taken from the baseline scenario, that is $N = 10$, $\alpha = 1$, $c = 0.33$. Other than in the analysis of illustrative scenarios, the optimal profile $\sigma(s^*, l^*)$ is now computed separately for every point in the parameter space, including for different levels of τ. Both sanctioning time s and cut-off level l can therefore differ across conditions, reflecting the adaptation of the optimal sanctioning profile to the changing requirements for deterrence and lenience as uncertainty and number of isolates change.

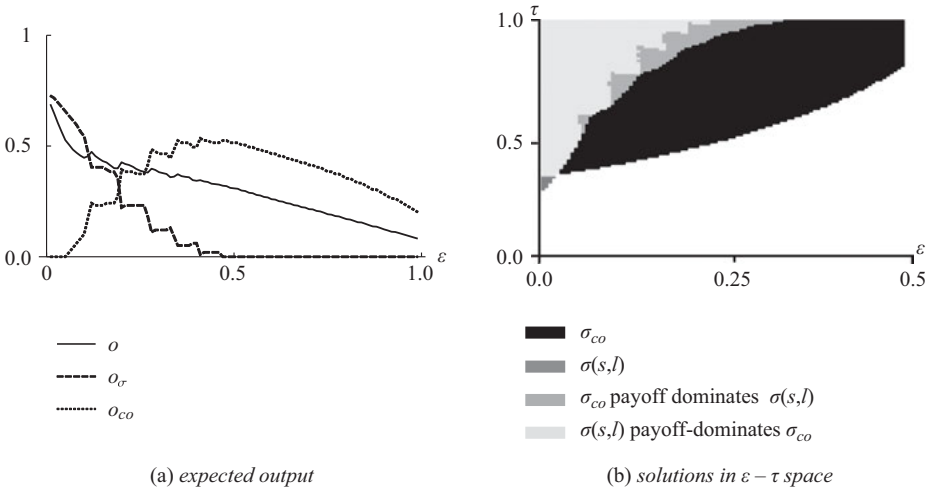
The share of the interval [0,1] of possible values of τ for which a particular type of strategy-profile is the solution of the game will also be used to obtain a coarse-grained indicator of the level of expected group-output o per round at a particular point in the parameter space. More precisely, for every level of τ, the corresponding solution of the game is identified and the related expected group-output $o(\tau)$ per round is computed. Overall expected group-output o at (α, c, N, N_c, ε) is then computed as average across the entire range of τ. For convenience the expected output is linearly rescaled to [0,1]. Details of the method how $o(\tau)$ is computed for a given profile $\sigma(s,l)$ can be found in Flache (2002). When the solution is the optimal $\sigma(s,l)$ profile, expected output is obtained as the probability that a particular round of the game falls within a normal period, multiplied with the rescaled expected group-output $(1 - \varepsilon)$ in a normal round. Finally, if the solution is σ_{co}, expected output is taken as $(1 - \varepsilon)N_c/N$.

To distinguish between output based on universal conditional cooperation and output based on core-only conditional cooperation, I also adapt the method described

above to obtain two additional output indicators. These indicate expected outputs separately computed for the fraction of the range of τ in which $\sigma(s,l)$ is s.p.e., and in which σ_{co} is s.p.e. These indicators are denoted o_{σ} and o_{co}, respectively.

5.3.2.2 Comparative statics for the effects of uncertainty ε

Figure 5.3 shows how the output indicators as well as the distribution of solutions in the τ-space change when uncertainty increases from $\varepsilon = 0$ to $\varepsilon = 0.5$, given $N = 10$, $\alpha = 1$, $c = 0.33$ and $N_i = 2$. More precisely, for this analysis both ε and τ are varied across 100 equidistant steps.



(a) *expected output* (b) *solutions in ε – τ space*

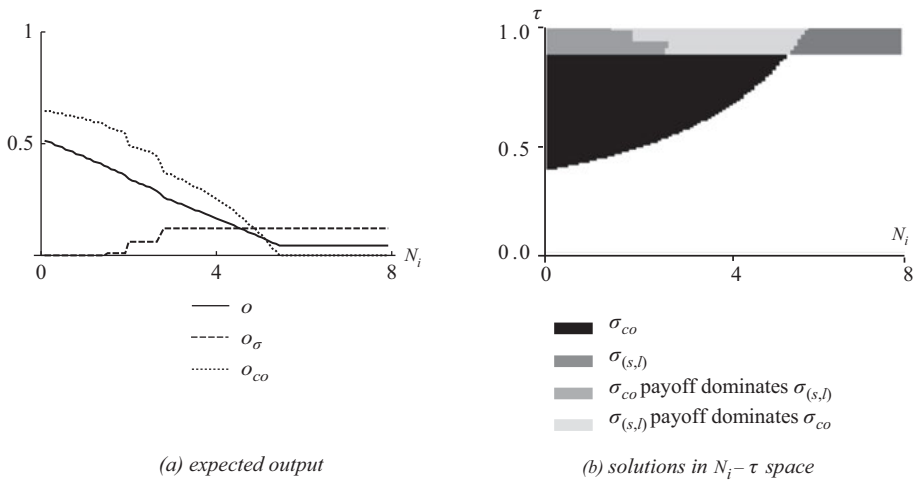
Figure 5.3: Effect of uncertainty ε on indicators of expected output and distribution of solutions in $\varepsilon - \tau$ parameter space. Baseline scenario: $N = 10$; $\alpha = 1$; $c = 0.33$; $N_i = 2$.

Figure 5.3b shows that at all levels of ε, any form of conditional cooperation is sustainable only when τ exceeds a critical threshold that becomes increasingly restrictive as ε increases. The figure also shows that except for very low levels of uncertainty, the conditions under which σ_{co} is the solution of the game are less restrictive than are the conditions for $\sigma(s,l)$. For approximately $0.02 < \varepsilon < 0.3$, $\sigma(s, l)$ is s.p.e. in an intermediate range of τ, but in parts of this region it is still payoff-dominated by σ_{co}. The erratic shape of those areas in Figure 5.3b is due to the discrete structure of the strategy space for $\sigma(s,l)$. For the uncertainty range of $0.02 < \varepsilon < 0.3$, the profile $\sigma(s, l)$ becomes the payoff-dominant solution only when τ exceeds an even higher threshold. From about $\varepsilon = 0.3$ on, $\sigma(s,l)$ can no longer be sustained for any level of τ, while this is still possible for σ_{co}. These results confirm the intuition that universal conditional cooperation is

less robust against uncertainty than core-only conditional cooperation. Figure 5.3a charts how this affects expected outputs o, o_σ and o_{co}. As uncertainty increases, expected output drops both overall and for o_σ. The output indicator o_{co} shows a different pattern. It first increases, then decreases in uncertainty. The reason is that starting from zero uncertainty, more uncertainty quickly widens the range of conditions under which core-only cooperation is the only individually rational strategy profile. Only beyond approximately $\varepsilon = 0.2$, efficiency losses from uncertainty also noticeably affect the expected output o_{co} that is sustained by the core-only profile.

5.3.2.3 Comparative statics for the effects of the number of isolates N_i

Figure 5.4 shows how changes in the number of isolates between $N_i = 0$ and $N_i = 8$ affect the distribution of solutions and expected output for the baseline-scenario $N = 10$, $\alpha = 1$, $c = 0.33$ and $\varepsilon = 0.2$. Both N_i and τ are varied across 100 equidistant steps in this analysis. While the number of isolates is conceptually a variable with only integer values, mathematically the corresponding indicators can also be computed for non-integer N_i . Quasi-continuous N_i was therefore used here to obtain a smooth representation of results.



(a) expected output *(b) solutions in $N_i - \tau$ space*

Figure 5.4: Effect of number of isolates N_i on indicators of expected output and distribution of solutions in $N_i - \tau$ parameter space. Baseline scenario: $N = 10$; $\alpha = 1$; $c = 0.33$; $\varepsilon = 0.2$.

Figure 5.4b shows that the range of conditions under which universal conditional cooperation $\sigma(s,l)$ yields an individually rational strategy-profile is not affected by N_i. In the baseline-scenario, τ needs to exceed approximately $\tau = 0.86$ to sustain

cooperation based on $\sigma(s,l)$, regardless of the value of N_i. The number of isolates has no effect on the rationality condition for $\sigma(s,l)$. In this profile, group members condition sanctioning only on the observed group-output. This implies that the sanctioning profile needed to keep a group member in line is the same for both core-members and isolates. Moreover, as long as the group contains at least one isolate there is also no other symmetric as well as efficient profile that guarantees that all group members including the isolate cooperate conditionally. Figure 5.4b demonstrates furthermore that the conditions under which σ_{co} is individually rational are less restrictive than those for $\sigma(s,l)$ as long as $N_i \leq 5$, but unlike for $\sigma(s,l)$ this range shrinks if a group contains more isolates. This happens because more free-riding isolates reduce the long-term benefits core-members can obtain under σ_{co} relative to the short-term gains from unilateral defection. Beyond $N_i = 6$, core-only conditional cooperation becomes unsustainable at all levels of τ. The figure further shows a region of conditions under which universal conditional cooperation payoff-dominates core-only cooperation, while both are individually rational. This happens when N_i is in an intermediate range (between about 2 and 6) and τ exceeds approximately $\tau = 0.86$. Above this level of τ it can be observed that the payoff-dominance relation between the two profiles is reversed for lower levels of N_i, because here efficiency losses under σ_{co} are relatively small compared to efficiency losses from sanctioning periods under $\sigma(s,l)$. At higher levels of N_i, σ_{co} is no longer sustainable and $\sigma(s,l)$ becomes the unique conditionally cooperative profile that is individually rational.

Figure 5.4a shows how the negative impact of N_i on core-only cooperation translates into a negative impact on the expected output a group with a core-periphery structure can achieve. While larger N_i comes with higher o_σ from about $N_i = 1.5$ on, the contribution that universal conditional cooperation can make to group-output is too small to compensate for the output-losses that core-only cooperation suffers from an increasing number of free-riding isolates.

All in all, Figure 5.4 confirms the intuition that a larger number of isolates makes core-only cooperation harder to attain and less efficient, such that universal conditional cooperation can become the superior solution given that group members are sufficiently interested in future outcomes.

5.4 Discussion and conclusion

Conditional cooperation can be a powerful endogenous solution to the problem of collective action if interactions within a group are "temporally embedded" (Batenburg et al. 2003). However, as several authors have pointed out, the level of provision of a collective good that can be sustained by endogenous conditional cooperation may be severely reduced if uncertainty blurs the link between the observed provision level of

a collective good and the underlying individual contribution decisions. Moreover, such uncertainty may also considerably reduce the range of conditions under which endogenous conditional cooperation is feasible at all for individually rational actors. In this case embeddedness in social networks can be an important condition that stabilizes conditional cooperation. Following up on Raub & Weesie's (1990) analysis, a large number of theoretical and empirical papers demonstrated that monitoring and control via social network relations can safeguard conditional cooperation especially for two-party cooperation problems, even when information about the degree of cooperation at the collective level is unreliable or not available.

In this paper I have proposed a first step towards integrating monitoring via direct social ties into a model of conditional cooperation in collective-good production under uncertainty. The model proposed here points to a dilemma that may occur especially in groups with a core-periphery structure. Members embedded in the densely connected core of a wider network face closer monitoring and can be more effectively sanctioned under conditional cooperation than members in peripheral positions in the network. To keep the latter in line, harsher sanctioning regimes must be imposed but these entail under uncertainty the possible cost of frequent 'erroneous sanctioning' which is detrimental for all group members. Thus, even for core-members themselves it can be a more attractive rational solution to adopt a cooperation norm that enforces contributions from core-members only and tolerates free-riding from peripheral members of the group. Yet, such an outcome can be deemed unfair, which imposes another social cost on the group.

The analysis presented here may point to testable hypotheses for empirical studies that aim to explain differences in cooperation in collective action between otherwise comparable groups that are different in their network structures. One example are organizations, like university departments, which can differ from each other in the relative number of full-time and part-time employees. Another are voluntary community-based initiatives, like local renewable energy initiatives, some of which operate in cohesive rural communities and some in more sparsely connected urban neighborhoods. A complication in formulating such hypotheses is that in a concrete collective good problem there are many other characteristics of actors that may affect contribution to a collective effort and are correlated at the same time with actors' position in a network. For example, members of a team of frontrunners in a community-energy initiative are likely to attach more value to the collective good of protection of the environment than community members who do not belong to this core-team, which by itself may explain why they would be more willing to contribute. However, the model proposed here also points to potentially unique testable implications. For example, it suggests that front-runners in a community energy initiative might be more willing to accept low levels of contributions from other community members without trying to respond with sanctions, the larger the core-team is relative to the community as a whole and the less easy it is for members of the core team to observe the reasons why others fail to contribute. At the

same time, their own level of contribution should not be reduced by those conditions. However, before empirical applications can be seriously addressed, a number of strong simplifications and potential limitations of the analysis presented here require careful inspection.

A first possible limitation is that I have focused only on a limited set of possible conditionally cooperative strategies. Obviously other and more complex conditionally cooperative strategy-profiles can be constructed. Especially profiles are of interest that combine the sanctioning-threat that core-only cooperation imposes on core-members with more lenient yet demanding norms for contributions by isolates. Such profiles could be more efficient than σ_{co} as well as less vulnerable to uncertainty than $\sigma(s, l)$. However, while this possibility cannot be excluded it should also be pointed out that such profiles can lead to prohibitively complex coordination problems as they involve different s and l levels for the different types of actors, as well as different degrees of contribution and sanctioning from core-members and isolates. This higher coordination complexity reduces the empirical plausibility of such solutions.

A second strong simplification I have adopted is the maximally simple network structure that distinguishes only two types of network positions, core-members and isolates. In real groups facing collective good problems networks are more heterogeneous. The analysis proposed here could be extended to such settings with the assumption that core-members will always defect if they observe in their personal network at least one verifiable defection. In this case, a defection of a core-member would trigger off a 'wave' of defection spreading through the network. Intuitively, this implies that members who are more closely connected also face stronger incentives to cooperate conditionally, because their defection leads to a faster and possibly more comprehensive breakdown of cooperation than the defection of more peripheral or even isolated network-members. On a qualitative level this retains the prediction of my simpler model that conditions for universal conditional cooperation are more restrictive in less densely connected networks, similar to results obtained in the research on effects of network embeddedness on cooperation in two-party interactions.

A third simplification is that my model abstracts from other differences between group members than their position in the network. For many settings it is plausible that that the network position is correlated with other differences between group members. One example is an organization in which members with longer tenure both face a darker shadow of the future and hold more central network positions than temporary workers. Another is a voluntary association, in which members with a larger interest in the collective good are also more active and therefore form part of the core of a network of activists. Moreover, mutual sanctioning and control among core-members can be expected to be even more effective if it is taken into account that network ties serve not only as channels for monitoring but also for imposing direct positive or negative sanctions upon other network-members (Coleman

1990; Fehr and Gächter 2002; Flache 1996; Flache et al. 2017). As long as more central network positions come with stronger incentives to contribute to the collective good (as in these examples), additional heterogeneity would not alter the proposition of a rational exploitation of the core by the periphery. However, more central network positions may also help core-members to enforce peripheral members' contributions more effectively. One possibility is that core-members hold positions with more formal sanctioning power in an organization, another that they use their social ties to coordinate effective monitoring and sanctioning efforts directed towards peripheral members (Coleman 1990). In addition, especially in organizations peripheral workers with temporary or part-time contracts may have strong incentives to contribute to collective organizational goods because this can enhance their prospect of acquiring tenure or other future rewards from the employer (cf. Lambooij, Flache, and Siegers 2009).

Fourth, a strong simplification of the model developed here is that networks are assumed to be static. In several of the examples discussed above, it would be possible that members of a group strategically change their network ties, for example in order to increase their possibility to observe others' contributions. Or, a principal could change organizational structures to reduce the isolation of peripheral members aiming to thereby facilitate endogenous cooperation. The very structures that inhibit conditional cooperation according to the model proposed here, could also be the outcome of processes of endogenous strategic network formation in which cooperators try to evade interactions with defectors (Sohn, Choi and Ahn 2019), or in which defectors try to insulate themselves from outside pressures to contribute (Takács, Janky and Flache 2008).

A last complication worth considering is the possibility of ostracizing isolates who free-ride (Hirshleifer and Rasmusen 1989). After all, under a core-only norm these members add nothing to the group-output. While this may be a readily available solution in a group that focuses only on one collective good to produce, in many empirical situations peripheral group members may make other valuable contributions that are unrelated to a specific collective good. Think for example of a university department that desperately needs temporarily employed teachers to offer all the courses needed, but also would like all employees to contribute to more-or-less voluntary collective activities such as organizing departmental colloquia or attending staff-meetings. It is quite plausible that in such cases norms emerge that demand perpetual contribution from tenured staff but tolerate some level of free-riding from less well connected temporary members of the organization.

Despite the limitations of the simple model I have proposed here, I hope to have demonstrated that studying how networks may sustain conditional cooperation through improved possibilities for monitoring is potentially a fruitful source of inspiration not only for the area of trust in two-party relations, but also for the study of cooperation in the production of collective goods.

References

Axelrod, Robert. 1984. *The Evolution of Cooperation*. New York: Basic Books.

Batenburg, Ronald S., Werner Raub and Chris Snijders. 2003. "Contacts and Contracts: Dyadic Embeddedness and the Contractual Behavior of Firms." *Research in the Sociology of Organizations* 20: 135–88.

Bednar, Jenna. 2006. "Is Full Compliance Possible?: Conditions for Shirking with Imperfect Monitoring and Continuous Action Spaces." *Journal of Theoretical Politics* 18 (3): 347–75.

Bendor, Jonathan and Dilip Mookherjee. 1987. "Institutional Structure and the Logic of Ongoing Collective Action." *American Political Science Review* 81 (1): 129–54.

Bendor, Jonathan and Dilip Mookherjee. 1990. "Norms, Third-Party Sanctions, and Cooperation." *Journal of Law, Economics, and Organization* 6 (1): 33–63.

Bendor, Jonathan, Roderick M. Kramer and Suzanne Stout. 1991. "When in Doubt . . . : Cooperation in a Noisy Prisoner's Dilemma." *Journal of Conflict Resolution* 35 (4): 691–720.

Buskens, Vincent. 2002. *Social Networks and Trust*. Dordrecht: Kluwer Academic Publishers.

Buskens, Vincent and Werner Raub. 2002. "Embedded Trust: Control and Learning." Pp. 167–202. In *Group Cohesion, Trust and Solidarity. Advances in Group Processes*, ed. Shane R. Thye and Edward J. Lawler. Amsterdam: Elsevier.

Chwe, Michael S-Y. 1999. "Structure and Strategy in Collective Action." *American Journal of Sociology* 105 (1): 128–56.

Coleman, James S. 1990. *Foundations of Social Theory*. Cambridge, Ma.: Harvard University Press.

Dijkstra, Jacob and Marcel A.L.M. van Assen. 2013. "Network public goods with asymmetric information about cooperation preferences and network degree." *Social Networks* 35 (4): 573–582.

Fatas, Enrique, Miguel A. Meléndez-Jiménez, and Hector Solaz. 2010. "An experimental analysis of team production in Networks." *Experimental Economics* 13 (4): 399–411.

Fatas, Enrique, Miguel A. Meléndez-Jiménez, Antonio Morales and Hector Solaz. 2015. "Public Goods and Decay in Networks." *SERIEs: Journal of the Spanish Economic Association* 6 (1): 73–90.

Fehr, Ernst and Simon Gächter. 2002. "Altruistic Punishment in Humans." *Nature* 415 (6868): 137–40.

Flache, Andreas. 1996. *The Double Edge of Networks. An Analysis of the Effect of Informal Networks on Cooperation in Social Dilemmas*. Amsterdam: Thesis Publishers.

Flache, Andreas. 2002. "The Rational Weakness of Strong Ties: Collective Action Failure in a Highly Cohesive Group of Rational Agents." *Journal of Mathematical Sociology* 26: 189–216.

Flache, Andreas. 2004. "How May Virtual Communication Shape Cooperation in a Work Team? A Formal Model Based on Social Exchange Theory." *Analyse & Kritik* 26 (1): 258–78.

Flache, Andreas, Dieko Bakker, Michael Mäs and Jacob Dijkstra. 2017. "The Double Edge of Counter-Sanctions. Is Peer Sanctioning Robust to Counter-Punishment but Vulnerable to Counter-Reward?" Pp. 280–301 in *Social Dilemmas, Institutions, and the Evolution of Cooperation*, ed. by Ben Jann and Wojtek Przepiorka. Berlin: De Gruyter Oldenbourg.

Flache, Andreas, Michael W. Macy and Werner Raub. 2000. "Do Company Towns Solve Free Rider Problems? A Sensitivity Analysis of a Rational-Choice Explanation." Pp. 123–125 (summary accompanied by full paper on CD-ROM, 36 pp.). In: *The Management of Durable Relations: Theoretical and Empirical Models for Households and Organisations*, ed. Jeroen Weesie and Werner Raub. Amsterdam: Thela Thesis.

Friedman, James W. 1971. "A Non-Cooperative Equilibrium for Supergames." *Review of Economic Studies* 38: 1–12.

Friedman, James W. 1986. *Game Theory with Applications to Economics*. New York: Oxford University Press.
Goedkoop, Fleur, Andreas Flache, and Jacob Dijkstra. 2017. "Participation within Community led Energy Projects: The Role of Social Networks. *Paper presented at Third European Conference on Social Networks* – EUSN 2017. Mainz, September 28, 2017.
Gould, Roger V. 1993. "Collective Action and Network Structure." *American Sociological Review* 58: 182–96.
Granovetter, M. 1985. "Economic Action and Social Structure : The Problem of Embeddedeness." *American Journal of Sociology* 91: 481–512.
Green, Edward J., and Robert H. Porter. 1984. "Noncooperative Collusion under Imperfect Price Information." *Econometrica* 52 (1): 87–100.
Hardin, Garrett. 1968. "The Tragedy of the Commons." *Science* 162: 1243–48.
Harsanyi, John C. 1977. *Rational Behavior and Bargaining Equilibrium in Games and Social Situations*. Cambridge: Cambridge University Press.
Hirshleifer, David, and Eric Rasmusen. 1989. "Cooperation in a Repeated Prisoners' Dilemma with Ostracism." *Journal of Economic Behavior and Organization* 12: 87–106.
Kollock, Peter. 1993. "'An Eye for an Eye Leaves Everyone Blind' : Cooperation and Accounting Systems." *American Sociological Review* 28: 768–86.
Kreps, David M. 1990. *A Course in Microeconomic Theory*. New York: Harvester.
Lambooij, Mattijs, Andreas Flache and Jacques Siegers. 2009. "Shadow of the Future, Risk Aversion, and Employee Cooperation." *Rationality and Society* 21 (3): 307–336.
Macy, Michael W. 1991. "Chains of Cooperation : Threshold Effects in Collective Action." *American Sociological Review* 56: 730–47.
Marwell, Gerald, and Pamela Oliver. 1993. *The Critical Mass in Collective Action. A Micro-Social Theory*. Cambridge (Mass.): Cambridge University Press.
Nikiforakis, Nikos. 2008. "Punishment and Counter-Punishment in Public Good Games: Can We Really Govern Ourselves?" *Journal of Public Economics* 92 (1): 91–112.
Nikiforakis, Nikos and Dirk Engelmann. 2011. "Altruistic Punishment and the Threat of Feuds." *Journal of Economic Behavior and Organization* 78 (3): 319–32.
Olson, M. 1965. *The Logic of Collective Action*. Cambridge, MA: Harvard University Press.
Raub, Werner. 1988. "Problematic Social Situations and the Large Number Dilemma." *Journal of Mathematical Sociology* 13 (4): 311–57.
Raub, Werner. 2017. *Rational Models*. Utrecht: Universiteit Utrecht.
Raub, Werner and Thomas Voss. 1986. "Conditions for Cooperation in Problematic Social Situations." In *Paradoxal Effects of Social Behavior. Essays in Honor of Anatol Rapoport*, edited by A Rapoport, Andreas Diekmann, and P Mitter, 85–103. Heidelberg: Physica-Verlag.
Raub, Werner and Jeroen Weesie. 1990. "Reputation and Efficiency in Social Interactions : An Example of Network Effects." *American Journal of Sociology* 96 (3): 626–54.
Raub, Werner and Jeroen Weesie. 2000. "The Management of Matches: A Research Program on Solidarity in Durable Social Relations." *Netherlands' Journal of Social Sciences : A Publication of the Netherlands' Sociological and Anthropological Society* 36 (1): 71–88.
Raub, Werner, and Vincent Buskens. 2008. "Theory and Empirical Research in Analytical Sociology: The Case of Cooperation in Problematic Social Situations." *Analyse & Kritik* 30 (2): 689–722.
Rooks, Gerrit, Werner Raub and Frits Tazelaar. 2006. "Ex Post Problems in Buyer-Supplier Transactions: Effects of Transaction Characteristics, Social Embeddedness, and Contractual Governance." *Journal of Management and Governance* 10: 239–76.
Selten, Reinhard. 1965. "Spieltheoretische Behandlung Eines Oligopolmodells Mit Nachfrageträgheit." *Zeitschrift Für Die Gesamte Staatswissenschaft* 121: 301–24.

Sohn, Yunkyu, Jung-Kyoo Choi and Toh-Kyeong Ahn. 2019. "Core–Periphery Segregation in Evolving Prisoner's Dilemma Networks". *Journal of Complex Networks* 8 (1): cnz021, https://doi.org/10.1093/comnet/cnz021.

Spagnolo, G. 1999. "Social Relations and Cooperation in Organizations." *Journal of Economic Behavior & Organization* 38: 1–25.

Takács, Karoly, Béla Janky and Andreas Flache. 2008. "Collective Action and Network Change." *Social Networks* 30 (3):177–189.

Taylor, Michael. 1976. *Anarchy and Cooperation*. London UK: Wiley & Sons.

Taylor, Michael. 1987. *The Possibility of Cooperation*. Cambridge: Cambridge University Press.

Wu, Jianzhong and Robert Axelrod. 1995. "How to Cope with Noise in the Iterated Prisoner's Dilemma. *Journal of Conflict Resolution* 39 (1): 183–189.

Wolitzky, Alexander. 2013. "Cooperation with Network Monitoring." *Review of Economic Studies* 80 (1): 395–427.

Siegwart Lindenberg, Rafael Wittek and Francesca Giardini

6 Reputation Effects, Embeddedness, and Granovetter's Error

Abstract: Reputation effects are crucial for social life. There has been important work done in the social sciences on this topic and Raub's contribution has been widely recognized. It builds on Granovetter's seminal work on embeddedness. However, Raub's contribution is unnecessarily limited by the fact that he copied Granovetter's error by assuming that all we need for dealing with reputation effects is attention to social structure (in the sense of networks) and to rational choice as a theory about actors. In our contribution, we argue that if reputation effects in the moral domain (compared to reputation effects in the domain of competence) work properly they inform people about the salience of overarching goals, including the very goal to follow normative obligations. To understand the conditions under which this happens necessitates attention to normative embeddedness, to normative heterogeneity, to structural features beyond networks (ingroup/outgroup dynamics and power differences), and to the mechanisms that govern the dynamics of overarching goals. This requires a serious correction of Granovetter's error, by approaching reputation effects in the moral domain on the basis of microfoundations that can deal with the interdependence between psychological processes and social structure.

6.1 Introduction

Reputation effects (especially those in the moral domain) have been heralded as one of the pillars of human cooperation. Milinksi (2016) even calls reputation "a universal currency for human cooperation." Reputation effects can make individually costly but socially beneficial behaviors more likely. Yet, it is by no means clear how reputation effects work and when they might fail to occur. For example, according witness testimony, Harvey Weinstein was sexually abusing women for years, without reputation effects. And then, all of a sudden, it was a scandal. Rebecca Traister (Traister 2017) asked in an article in The Cut in October 2017 why the Weinstein allegations did not make a stir much earlier. "The history of allegations has been an almost wholly open secret [. . .] and yet somehow ignored, allowed to pass, unconsidered." Why did similar allegations hit the press about Trump and yet, nothing much happed? Why did the reputation effect not work properly in either of these cases? Journalists

Siegwart Lindenberg, Department of Sociology (ICS), University of Groningen; Tilburg Institute for Behavioral Economics Research, Tilburg University
Rafael Wittek, Francesca Giardini, Department of Sociology (ICS), University of Groningen

https://doi.org/10.1515/9783110647495-006

have good guesses, but our scientific grasp of these dynamics clearly needs help, by considering more closely the conditions under which reputation effects are likely and conditions under which they are unlikely to occur.

Seemingly, the widespread and simple idea about reputation effects is that because people care about what is said about them by others, they will behave in such a way that nothing bad is said about them. In addition, they base decisions about their own behavior on what they have seen or heard about the other's behavior in the past (for example, they do or don't trust the other's promise). In the literature, the behaviors that are allegedly regulated by reputation effects are often lumped together under the term "cooperation," meaning basically what Granovetter (1985) called "orderly transactions," including trustworthiness and prosocial behavior. Gintis and Fehr (2012: 28) remark that this approach to cooperation is by now quite standard in economics and game theory, with the famous combination of rational choice and assumed self-interest: People engage in orderly transactions not because they have a feeling of obligation to do so, but because they want to avoid getting a bad reputation. The dynamics of social obligation and shared social norms plays no role in this approach, leaving an open goal for sociologists to at least try to kick in the ball. Yet, counter to what one might expect, the sociological research that picked up reputation effects ever since Granovetter's (1985) work on embeddedness of human transactions in dyadic relationships and social networks, more or less ignored the importance of being embedded in shared norms and the mechanisms involved in their workings (for example how norms influence behavior differently in when ingroup/outgroup dynamics are at play or when there are great power asymmetries).

Werner Raub is one of the sociologists who, following Granovetter, have prominently contributed to the reputation and embeddedness research. But he followed Granovetter and saw the embeddedness effects based on the assumption of rational choice and self-interest as more parsimonious (and thus preferable) explanations of cooperation and trust than explanations that are based on moral commitment (that is a feeling of obligation to follow social norms). This pitting of rational choice against "normative" or "moral" explanations of cooperation and trust is in our view based on an erroneous contrast between rational choice and normative explanations, and it unnecessarily limits what can be learned from research on reputation effects. But before we expound on the reasons why we consider this position flawed, we will briefly explain the considerable role Granovetter himself played in this development.

6.2 Granovetter's error

In 1985, Granovetter published a by now much cited article about "embeddedness". The general message was that two prominent views on how "orderly transactions" (in business transactions and life in general) are brought about, are fundamentally

flawed. The classical view in sociology of orderly transaction as being brought about by internalized norms and values, and the classical and neoclassical view in economics of orderly transaction as being brought about by clever institutions and self-interest are both atomistic, neglecting the impact of social structure. Dyadic relations are embedded in relations with others and between others and this embedding has great influence on behavior, including behavior in business transactions. "Actors do not behave or decide as atoms outside a social context, nor do they adhere slavishly to a script written for them by the particular intersection of social categories that they happen to occupy. Their attempts at purposive action are instead embedded in concrete, ongoing systems of social relations" (Granovetter, 1985:487). The sociological program to be pursued thus should be to study the influence of embeddedness on the likelihood that transactions are orderly. For sure, Granovetter pointed to an important shortcoming of economic explanations of orderly transactions, and Werner Raub followed Granovetter's program, made it his own and pursued it in a very systematic way. He focused particularly on trust problems, that is, on situations in which lack of trust or abuse of trust would imply results that are inferior to the social beneficial outcome (Raub and Buskens 2008).[1] He studied various kinds of embeddedness that affect trust: dyadic (that is the same dyad embedded by shared past or future), network, and institutional. In the course of realizing this program with colleagues and students, he investigated relational and structural control effects (based on the ability to sanction and reward via future interactions or via spreading information) and learning effects (based on experience or information from others). In addition, he investigated how institutions (for example contract law or an eBay feedback forum) positively or negatively affect relational and/or network embeddedness.[2] Yet, with all this positive development, there is also a highly problematic carry-over from Granovetter's embeddedness program. It can be called "Granovetter's error".

Granovetter (1985) introduced his embeddedness argument together with a heuristic that indicated in what direction progress may be possible: focus on the effects of structure and assume rational choice as a working hypothesis to deduce that people rationally consider structural effects (such as reputation effect) for their action. He admits, that rational choice theory uses naïve psychology, but

> the notion that rational choice is derailed by social influences has long discouraged detailed sociological analysis of economic life and led revisionist economists to reform economic theory by focusing on its naive psychology. My claim here is that however naive that psychology may be, this is not where the main difficulty lies – it is rather in the neglect of social structure.
>
> (Granovetter 1985: 506)

1 This too follows Granovetter's (1985) lead, who focused on the effect of embeddedness on "trust and malfeasance".

2 Even though we maintain that too little can be learned from this approach about reputation effects "in the wild", we have no space to go into a discussion of the achievements of Raub's research program.

Granovetter (1985: 505) speaks of "psychological revisionism" which in his view is something one should stay away from, "an attempt to reform economic theory by abandoning an absolute assumption of rational decision making". In the embeddedness approach, once we work with the behavioral "working hypothesis" of rational choice, "the details of social structure will determine which is found." (Granovetter 1985: 493).

Of course, Granovetter's approach is much to be preferred over and above a pure structuralist approach that does not even allow any theory of action to help explain structural effects (see Lindenberg 1995 for a critique of the pure structuralist approach). However, it is infected by a fundamental error: it makes the untenable assumption that structural effects are independent of the working of norms and the psychological mechanisms connected to the way they work. Yet, contrary to what would be implied by Granovetter's error (and often found in the economic literature, for example Fehr, Brown and Zehnder 2009), long-term business relationships are not likely the result of the ongoing working of reputation effects. Granovetter in fact does not keep to his own program and brings in norm-related psychological mechanisms through the back door. For example, in a sequel to his 1985 article, he explains why intimate relationships have different structural effects than relationship with strangers, maintaining that

> continuing economic relations become overlaid with social content that, apart from economic self-interest, carries strong expectations of trust and abstention from opportunism. That is, I may deal fairly with you not only because it is in my interest, or because I have assimilated your interest to my own (the approach of interdependent utility functions), but because we have been close for so long that we expect this of one another, and I would be mortified and distressed to have cheated you *even if you did not find out* (though all the more so if you did).
>
> (Granovetter 1992: 42)

It is difficult to see how the rational choice "working hypothesis" would explain how "expecting fairness of each other" can lead to being "mortified and distressed to have cheated" even if the other did not find out. Also, if it is difficult to judge whether or not somebody cheated or was simply unable to keep a promise (as often happens), what cues in the interaction would help indicate a mishap rather than willful cheating? How is sincerity socially communicated and accepted? (Lindenberg 2000). Granovetter offers no hint on what the role of norms in all this might be. In order to trace different effects of being embedded in different kinds of relationships, Granovetter needs to make use of mechanisms he "officially" deemed to be irrelevant and whose use belongs to "psychological revisionism." There is a price to be paid for this kind of "shadow methodology" because the "bootlegged" psychological mechanisms cannot be analyzed, critically evaluated and possibly adapted or corrected.[3]

3 The use of such a shadow methodology has also been observed for Durkheim who officially rejected using psychology but in practice made ample use of psychological mechanisms in his

They must remain ad hoc, in the vague sphere of "common knowledge". Granovetter explicitly appeals to common knowledge to bolster his ad hoc theory of intimate relations: "It would never occur to us to doubt this last point in more intimate relations, which make behavior more predictable and thus close off some of the fears that create difficulties among strangers." (Granovetter 1985: 490). But to be on the safe side with regard to the phenomena that can occur in intimate relationships, he insists that "in personal relations it is common knowledge that 'you always hurt the one you love'; that person's trust in you results in a position far more vulnerable than that of a stranger." (Granovetter 1985: 491). Now, he can use one "common knowledge" assumption when he needs the positive effect of intimate relationships and the opposite one, when he needs the negative effect. Too little is gained in terms of what we learn about the dynamics of embeddedness by using these kinds of ad hoc assumptions.

To his credit, even though he followed Granovetter's program, Raub steadfastly refused to also follow his shadow methodology. To the contrary, he sharpened Granovetter's rational choice "working hypothesis" by keeping as strictly to the assumptions of standard game theory as possible. "We assume a setting of strategically interdependent actors, which seems appropriate to the embeddedness argument in general and to the modeling of reputation effects in particular. Thus, to analyze rational behavior in such a setting, we have to use strong game-theoretic rationality assumptions." (Raub and Weesie 1990: 629). This has the advantage of allowing rigorous model building and rigorous deductions. Yet, the price is that ever more simplifying (and thus unrealistic) assumptions have to be made to allow such rigor. For example, in 1990, Raub and Weesie published a by now much cited article on reputation effects (Raub and Weesie 1990) in which they contrast three game-theoretic models of reputation effects: one with no embedding, one with perfect embedding, and one with imperfect embedding. The models are rigorous but require that structural embedding is always associated with actors *costlessly* getting information on *all* interactions of their partners. In addition, there is no allowance for mistakes or misunderstandings. If an act is taken to be a "defection", it will trigger retaliatory reactions throughout the network. While some game theorists make this the basis for the generation of trust (others trust their transaction partner because they reckon with their partner's fear of this retaliatory reaction, Guennif and Revest 2005), Raub and Weesie admit that this lack of allowance for mistakes or misunderstandings is an extremely unstable social system. They suggest a remedy that keeps entirely to the inner logic of their models. Rather than internalizing people's fear of mistake and misunderstandings, they argue that, to avoid such instability "requires that actors be informed not only on the interactions of their partners

explanations (see Lindenberg 1983). Homans (1964: 818) said of shadow methodologists jokingly "they keep psychological explanations under the table and bring them out furtively like a bottle of whiskey, for use when they really need help."

but also on the interactions of the partners of their partner." (Raub and Weesie 1990: 464). It is a nice example of how highly simplifying assumptions can drive one progressively further away from social reality and from the necessity to deal with the dynamics of norms. Following Granovetter's error even without shadow methodology and with mathematical rigor will thus also teach us too little about the dynamics of reputation and embeddedness.[4]

6.2.1 Reputation and the importance of norms

The fateful and, in our view misguided, juxtaposition of rational choice versus "normative" explanations on which Granovetter's error is based, is embraced by Raub even more willingly, as it echoes an earlier and even more extreme statement by Coleman (1964: 166f), that Raub frequently cites: "sociologists have characteristically taken as their starting point a social system in which norms exist, and individuals are largely governed by those norms. Such a strategy views norms as the governors of social behavior, and thus neatly bypasses the difficult problem that Hobbes posed [. . .] I will proceed in precisely the opposite fashion [. . .] I will make an opposite error, but one which may prove more fruitful [. . .] I will start with an image of man as wholly free: unsocialized, entirely self-interested, not constrained by norms of a system, but only rationally calculating to further his own self-interest." This leave little room indeed for considering the relevance of norms and normative obligations for reputation effects.

Raub and Weesie (1990: 629) define reputation as "a characteristic or an attribute ascribed to him by his partners." However, they fail to add that it is not just any characteristic or attribute. Reputation is a socially acquired *evaluative* opinion about a social actor regarding a *tendency* to act in a particular way (be that an individual, group, organization or country) (Giardini & Wittek, 2019; Weigelt and Camerer 1988). Reputation is similar to direct social control (say, that one wants to avoid disapproval) with regard to the importance of shared norms and standards. However, reputation is more than direct social control. It is about a good/bad continuum, and that requires shared standards of evaluation *and* it is about a presumed behavioral tendency, even though this judgment may be based on just one incident ("once a liar, always a liar"). Thus, reputation effects are about the assumed likelihood that a social actor behaves dishonestly, is helpful, is better than others in certain regards, keeps promises, is cooperative, is opportunistic, etc. It is also useful to stress that unless one deals explicitly with dyadic reputation effects, reputation refers to the evaluative opinions that are shared in a particular group. In

4 Incidentally, Corten et al. (2016), testing the Raub/Weesie model of reputation effects found that its predictions did not hold up. Unfortunately, Corten et al. did not see Granovetter's error at work.

this sense, one can also speak of the reach of reputation, depending on how inclusive the group is in which it is shared. In order to see the importance of the evaluative aspect of reputation also for Granovetter and Raub, it is useful to observe the language that is used by them when describing the behavior about which reputation effects emerge: In Granovetter's (1985) article on embeddedness, the term "opportunism" appears 17 times, and "malfeasance" 25 times. In the Raub and Weesie (1990) article, "opportunism" appears 15 times, "malfeasance" 11 times, and "defection" 62 times. It is not said but implied that there are shared standards and norms, and that people who violate these standards or norms (that is defect, are opportunistic, or knowingly commit a wrongful act) must fear for their reputation because others may infer an action tendency on the basis of this violation. People anticipate such loss of a good reputation and thus are more likely to behave "properly."

Simply assuming clashing preferences or goals is not enough for reputation effects. If preferences and personal goals are important for reputation effects, their influence derives from their link with norms. For example, if an overnight guest squeezes my toothpaste in the middle and I happen to be somebody who gets irritated by that, it makes no sense to say that this is an example of how a bad reputation is being built. Even in the case of a purely dyadic reputation effect, my friend would have to know that squeezing the toothpaste in the middle is negatively evaluated by me, and I would have to know that he knows that, as he would have to know that I know he knows. In short, for reputation effects to occur in this situation, I would have to be thought of by my friend as somebody who willfully irritates a friend. What would be violated in this case is a friendship norm via my willingness to let personal preferences (for squeezing the toothpaste in the middle) prevail over behavior appropriate to a friendship relation.

People can have personal norms, but that is also not enough for reputation effects. It is the fact that standards and norms are shared that allows people to abstain from a certain behavior for fear of losing their good reputation. For example, if there is no shared norm about going to church on Sunday, people will neither lose their good reputation by not going, nor would they go to church for fear of losing their good reputation. Reputation effects thus necessitate normative embeddedness, that is shared standards and norms. The argument that norms cannot be taken as given, that they must be "endogenized" and explained by equilibrium behavior in repeated games (see Buskens and Raub 2013), is misplaced in this context, as reputation effects presuppose the existence of shared norms rather than the other way around. Norms do not emerge to make reputation effects possible, but reputation effects go piggyback on shared norms, that is they are made possible and derive their usefulness on the basis of shared norms. When it comes to examples of how reputation effects work, Raub and Weesie (1990: 642) don't hesitate to cite as prime example the case of diamond merchants (also used by Granovetter) that clearly presupposes shared norms: "Like other densely knit networks of actors,

they generate clearly defined standards of behavior easily policed by the quick spread of information on instances of malfeasance."

The requirement that reputation effects necessitate normative embedding is only half of the story. We also have to deal with the question why people would make judgments about "behavioral tendencies" on the basis of observed behavior. Is it simply that there are types of people and even one or a few behaviors signal a type? Or are these judgments more complex? This too is a question that needs answers.

Why is it important to make the dynamics of normative embeddedness and behavioral tendency judgments explicit? The quick answer to this question is: because norms and judgments about behavioral tendencies in dealing with norms affect virtually all aspects of the dynamics of reputation, be that the question what can be subject to reputation effects, where reputation effects can be expected, why these effects may vary in strength, when and why people would pass on reputation-relevant information, or how power asymmetries affect what is and is not damaging for one's reputation. It is exactly because reputation effects are such an important part of (self)controlling behavior that we need to be able to answer such questions. For answering any one of these questions, one has to consider psychological aspects connected with the dynamics of normative embeddedness, the actor who may or may not anticipate reputational effects, the observer who may or may not react to transgressions, and the transmitter who may or may not pass on information that could affects one's reputation.[5]

6.3 The effects of heterogeneity on normative embeddedness

Normative embeddedness that is needed for reputational effects depends on normative homogeneity and thus is not simply embeddedness in social relationships. This is far from being trivial. Many of the examples used in the literature about reputation effects are from anthropological research on so-called traditional societies, or from experiments with subjects from homogeneous populations (for example Giardini and Conte 2012; Gintis, Smith and Bowles 2001). In these examples, shared norms and standards are more or less taken for granted and not included in the analysis. But can one simply ignore normative embeddedness in our societies as well? As populations become more heterogeneous with regard to social norms, reputation effects are likely to become more restricted to normatively homogeneous

5 Because of restrictions of space, we follow Granovetter as well as Raub in focusing mostly on the dynamics concerning "bad" reputation.

local "pockets", and even friendship networks may be too heterogeneous for many reputation effects (say with regard to tax evasion, or honesty in business deals, or duties as a parent). As a result, reputation effects become less of a regulatory force for interpersonal behavior, with some authors even complaining about an increase in "cheating culture" (for example Callahan 2004). Instead, widespread reputation effects become more restricted to useful but quite mundane forms of evaluation (such as the quality of food in a restaurant, quality of online services, or e-commerce rating systems that solve the very trust problems created by the internet itself) because for them there is still a widespread consensus on standards. But even these reputational mechanisms are quite complicated. "Strong reciprocity" (that is the assumed predisposition to unconditionally reward or punish one's interaction partner's cooperation or defection, see Diekmann et al. 2014), is often invoked for such mechanisms. However, what is less often considered is that strong reciprocity also works against reputation effects by blocking feedback. For example, a study of reputation effects for eBay found that buyers are reluctant to give negative feedback for fear of counter punishment from the seller (Li 2010). In short, the mundane forms of reputation effects require a good deal of psychological finesse to work properly (see for example Bolton, Greiner and Ockenfels 2013 and 2018). In addition to the mundane effects, many reputation effects occur for the extreme forms of deviant behavior (such as pedophilia, sexual abuse of women, downright corruption). And even here, important psychological mechanisms are at work, drawing our attention to the role of "publicizers" (see the section on "transmitters" below). This leaves large gap of behavior that is not well covered. For example, for businesses, so-called reputation effects are increasingly "managed effects" based on corporate identity, certification contests, brand recognition, and quality perceptions realized more through public relations campaigns than observed past behaviors (see for example Gray Balmer 1998 or Balmer and Gray 1999). Alternatively, for business transactions, care for establishing and maintaining long-term relationships through bonding, relational contracting, and relational governance reduces the reliance on "fear" factors such as hostages and reputational effects inside organizations (Birkinshaw, Foss and Lindenberg 2014) and between organizations (Dyer and Singh 1998). Transgressions regarding the relevant norms involved (such as breaking trust) then may be fatal for the bonding of the relationship. For example, investigating interfirm networks in the apparel industry, Uzzi (1997: 59) found that "if the strong assumptions of trust and cooperation are exploited in embedded ties, vendettas and endless feuds can arise." This is likely to help the transacting parties from breaching trust. But it is not a reputation effect.

For shorter term relationships, a good part of what reputation effects could have done is taken over by institutional sources of evaluative information (such as information on creditworthiness, on certifications, on "good behavior" certificates, on rankings, ratings and comparative sites). Not surprisingly, where reputation effects regarding more complex evaluation questions of products and practices in

informal groupings are operative, we see explicit work on normative embeddedness with shared norms and standards. Examples are occupational communities (for example Lawrence 1998) and the increasing number of online communities with a focus on community building, mutual support and the sharing and exchange of information (Preece, Maloney-Krichmar and Abras, 2003).

Even when norms are widely shared, their ability to create reputational effects is dampened in heterogeneous societies. For many groups, growing heterogeneity (religious and ethnic) makes those norms that are shared more abstract. This was already observed by Durkheim (1964 [1893]): Norms in heterogeneous societies "rule only the most general forms of conduct and rule them in a very general manner, saying what must be done, not how it must be done." The consequence is that, in any concrete action situation, the abstractness of norms leaves considerable moral "wiggle room" (Dana et al. 2007; Lindenberg 2008; Mazar, Amir and Ariely 2008; Spiekermann and Weiss 2016), making it less obvious whether or not people transgressed a norm. Given such wiggle room, the actor is less concerned about reputational effects of many of her actions, and the observer is less likely to draw reputational conclusions from observing these actions. In sum, heterogeneity creates "wiggle room" with regard to norm conformity, which, in turn, weakens reputational effects. But there is much more to be considered than this wiggle room. Psychological aspects need also to be explicitly considered for all major figures involved in reputation effects: the actor, the observer, and the transmitter.

6.4 The actor

For reputation effects to occur, actors need to be concerned about them, but when is this the case? The damaging effect of a reputation effect derives from the possible inference that the "deviant" behavior reveals a behavioral tendency. Why would an actor assume that others do not only judge his actions, but also infer a *behavioral tendency* on the basis of first or second-hand experiences of his actions? To answer this question, we have to have a closer look at the role goals play in this. People conform to norms and legitimate rules and standards for different reasons, depending on their goals. It may be prudent to do so; or it may feel good to do so; or it may be the right thing to do. Reasons for doing things cluster when they are generated by the same goal. Particularly interesting in this regard are three overarching goals because their salience can change from one situation to another (Lindenberg and Steg 2007, 2013). It is very likely that the actors know that their behavior says something about the salience of their overarching goal. For example, Hilbe, Hoffman and Nowak (2015) showed with cleverly designed experiments that people who deliberately ignore cues that might tell them whether or not defection might be profitable are more trusted. "Intuitively, by not looking at the payoffs, people indicate that they will not

be swayed by high temptations to defect, which makes them more attractive as interaction partners." (Hilbe, Hoffman and Nowak 2015: 458). Behaviors are signals about which overarching goal is salient (Lindenberg 2000).

There are three overarching goals: a gain goal (focused on increasing or maintaining one's resources, such as money); a hedonic goal (focused on improving or maintaining the way one feels); and a normative goal (focused on acting as a member of a collective, on acting appropriately with regard to the norms of this collective). For goals to become action-relevant, they have to be cognitively salient. At any given moment, one of the three overarching goals is likely to be the most salient and have the strongest influence on reasons for doing things, which, in turn are linked to cognitive and motivational processes such as what one pays attention to and what one ignores, what concepts and chunks of knowledge are being activated, what alternatives one considers, what information one is most sensitive about, and how one processes information. Social cues can greatly affect the salience of these overarching goals (Lindenberg 2012), so that social embeddedness (structure) and normative embeddedness (shared norms with more or less salient normative goals) interact. Let us make this more concrete by discussing each of the three overarching goals from the point of view of the actor. Even though in real life, all three goals are active to some degree at the same time (with one of them being most salient), we will ignore this mixed motive aspects for reasons of simplicity of exposition.

Gain goal. The gain goal becomes salient in situations in which there are clear opportunities for increasing one's resources, such as money or status, investment opportunities, competitive conditions, and "golden" opportunities for profit. When the gain goal is salient, the focus is on costs and benefits of a particular action in the medium and longer term, and thus a person is likely to also consider whether his or her action can have damaging reputational consequences. Most economists (as well as Granovetter and Raub) assume that people only have this overarching goal. With regard to reputation effects, the actor with a salient gain goal anticipates some future interaction or some indirect effects via gossiping. If he assumes that nobody would know about his action, the action would not be constrained by reputational concerns and only focused on gain. Thus, people with a salient gain goal act strategically, and for them reputational effects are part of the cost/benefit calculation. They will consider the likelihood that their actions will be observed or traceable, whether at least some of the potential observers will judge the actions evaluatively, and whether the observers could gain from passing on evaluative information to third parties who do care about conformity to norms and standards. Without it being mentioned that somebody has to care about norms and standards, this is also as Granovetter and Raub would assume it happens. But reputation effects only contribute to orderly transactions if they are truly informative about somebody's behavioral tendency. Strategic self-interest (a salient gain goal) can undermine reputation effects because actors with a salient gain goal are likely to also

consider how they could appear to conform to norms and standards, even when they don't conform. In other words, when at least some relevant others are assumed to care about this conformity, actors with a salient gain goal will put effort into managing the impression they make on others even when they transgress norms and standards with impunity (see for a vivid description Williams and Milton 2015). The better they succeed in doing so, the less "orderly transactions" are brought about by anticipated reputation effects.

What is the assumed naïve psychology of others in this regard? If observers assume the gain goal to be salient where a salient normative goal was socially expected (say, after observing the breach of a promise), they also assume that the gain goal will easily dominate the normative goal on other such occasions. This is how one observed behavior can lead to the inference of a behavioral tendency. In fact, the naïve psychology associates a behavioral tendency with who you are, with an identity, and it influences behavior. For example, "please don't be a cheater" works much better in preventing unethical behavior than "please don't cheat" (Bryan et al. 2012). The inference process is based on incongruence of expected and inferred salience of an overarching goal. For example, a salient gain goal would be expected when somebody negotiates a deal; a salient hedonic goal would be expected when celebrating a birthday; and a salient normative goal would be expected with regard to promises made. If breaching the promise would be profitable, observing that it is breached without excuse or explanation is taken as a sign that the gain goal becomes salient too easily in this situation for this particular person or company, when it should not become salient at all (Lindenberg 2000).[6] Institutions actually encourage the salience of these overarching goals in these different situations (Lindenberg 2017). This, then, is the likely mechanism behind the link between behavior as a signal and the behavioral tendency that is assumed on the basis of this signal. The identity inference is based on the assumed ease with which one of the three overarching goals can become salient in a particular situation. Types (say a "cheater", "a money-grabber", a "saint") are of course not only identity labels people assign to others, but they do exist as chronically salient hedonic, gain or normative goals in people. For understanding reputation effects, however, it is also important to remember that, despite the chronic salience, *intra*personal shifts in the salience of overarching goals are also possible for such types, when the situation is strong enough (see for example Pulford et al. 2016). In sum, it is the incongruence between expected and experienced salient goals in others that is at the basis of reputations. Such incongruence is much less likely and thus much more informative than congruence. For this reason, it is also said that "it takes a long time to build a reputation and only an incident to ruin it."

6 This would hold even if the deviance were not related to moral standards but to standards of competence, but in the following, we will not go further into reputation effects for (in)competence, because they are likely to depend also on conditions specific to standards of competence.

As we discussed already, reputation effects are not about clashing "private" preferences or goals. It is by now clear that reputation effects actually rest on the fact that there are different overarching goals and that observed behavior is used to form impressions about their salience. Could negative reputation effects occur if everybody assumed everybody else to have a salient gain goal (as in, say, highly corrupt circles)? Our answer is: no. Even though everybody might prefer others to keep promises, to help, to be honest etc., if nobody assumes others to have a salient normative goal, there would be no news value attached to not keeping one's promises, not helping, not being honest. At best, one could acquire a reputation for keeping one's promises, being honest, etc., but with the assumption by the researcher that everybody has a salient gain goal, even this positive reputation effect would not be possible. This points to the internal inconsistency of an approach that assumes a gain goal for everybody (that is rational choice with self-interest and common knowledge) and claims to explain reputational effects. In a world without different overarching goals, reputational effects in the moral domain would play no important role.

Hedonic goal. The hedonic goal becomes salient when people are exposed to situations that contain strong visceral or arousing stimuli or when people are tired. When the hedonic goal is salient (say, somebody is in a party mood, or feels very anxious), the focus is on feelings (short term) and thus also on behavior that feels good or improves the way one feels. One is sensitive to others' proximal affective reactions (such as anger or praise), but this is different from anticipated reputational effects, as the concern is with the impact of the others' reactions on the way one feels right now, rather than about other people's inferences about one's behavioral tendencies. Thus, with a high concentration of people with a salient hedonic goal (such as often in a beach resort), we would expect much mutual concern about others' affective reactions, but only a very limited occurrence of behavior that is controlled by anticipated reputation effects.

The hedonic goal plays also an important role in reactions to observing a violation of norms by others, because such a violation often increases the salience of the hedonic goal, that is, it increases the focus on how one feels. For example, moral outrage after observing moral transgressions is often a hedonic reaction rather than a normative one (O'Mara et al. 2011; Veldhuis et al. 2014). A focus on how one feels makes reaction to observing norm violations very dependent on factors that have little to do with the spread of veridical information about trustworthiness (Wetzer, Zeelenberg, and Pieters 2007). Feeling victimized my lead to taking revenge and telling others an exaggerated account of what happened, or it may heighten one's vigilance and fear of counter attacks, making one abstain from passing on negative information (see for example Bolton, Greiner and Ockenfels 2013). We will come back to this issue then dealing with the observer and transmitter.

Normative goal. When the normative goal is salient, the focus is on behaving appropriately, with a feeling of obligation to do so. In this case, one would behave appropriately even when one feels unobserved. The salience of the normative goal is strongly influenced by institutions, such as the rule of law (Lindenberg 2017). However, this does not mean that others play no role. There is an evolved sensitivity to social cues that affects the salience of the normative goal and creates what we call "cued" reputation effects (in contrast to "calculated" reputation effects that can occur with a salient gain goal).[7] With calculated reputation effects, the presence of others could make one conform to norms because one wants to avoid getting a bad reputation. In contrast, even though cued reputation effects may have evolutionary roots in calculated reputation effects (Hoffman, Yoeli, and Navarrete 2016), they are by now governed more or less automatically by cues emanating from others. People with a salient normative goal have been shown to be less calculating about reputation effects (Simpson and Willer 2008) and signaling that one is not calculative increases trustworthiness in others (Jordan et al. 2016). In this way, others become more calculative and less prosocial vis-à-vis somebody who has a bad reputation (Schilke and Cook 2015; Wedekind and Milinksi 2000).

The physical and even the sheer psychological presence of others who stand for shared norms increases the salience of the normative goal (that is of the collective orientation) and it activates anticipated shame and/or guilt about not acting appropriately (Engelmann, Herrmann and Tomasello, 2012; Shaw 2003a, 2003b; Wu, Balliet, and van Lange 2016). That this is not part of a cost/benefit calculation can be gleaned from the fact that the salience of the normative goal is strengthened already by pictorial representations of eyes gazing at an individual (Manesi, van Lange, and Pollet 2016). The salience of the normative goal is also very responsive to the observed respect by others for norms and standards (Lindenberg, Six and Keizer 2020). For example, in field experiments, it was shown that observed (dis) respect for norm A greatly increased (decreased) conformity to norm B in the observer (see Keizer, Lindenberg and Steg 2008, 2013). Normative and social embeddedness interact but cued reputation effects are not related to the anticipation of a possible flow of information through specific networks. Rather, cued reputational effects are linked to the presence (bodily or psychological) of generalized others and signs of their respect or disrespect for norms and standards. Thus, even though network structures may be important for cued reputation effects and for what information (if any) people will circulate, other structural features, such as ingroup/outgroup differences and power asymmetries, maybe at least as important.

7 In the literature, these two ways in which reputation effects restrain behavior are often confounded (for example Jordan et al. 2016).

6.5 Special effects for the actor: Ingroup/outgroup and power-asymmetry effects

Ingroup/outgroup effects. Human beings have evolved to function in so-called fusion/fission groups (Aureli et al. 2008). This means that they are cognitively and motivationally equipped to function in changing group constellations, such that being part of, say, a hunting party in the morning and part of an inclusive group (comprising various subgroups) for arranging a safe camp for the night. The collective to which the normative goal refers thus can change flexibly, depending on the circumstances. This ability also allows so-called ingroup/outgroup dynamics, in which the relevant collective for the normative goal explicitly excludes some other group (the outgroup). Intergroup conflict and competition will strengthen this "parochial" orientation of the normative goal (De Dreu et al., 2016; Wildschut and Insko, 2006). Thus, ingroup members expect each other to have a salient normative goal regarding the ingroup (and be rewarded with social approval and status for actively "proving" these orientations to their ingroup fellows, see De Dreu, Balliet, and Halevy 2014) and to have a salient gain goal or hedonic goal regarding the outgroup. Outgroup members are similarly expected to be parochial in their orientation. For an actor, this has important consequences regarding reputational effects: strong cued reputational effects for most ingroup members, strong calculated reputational effects for some ingroup members, and for all ingroup members no reputational effects vis-à-vis outgroup members. Thus, whatever makes groups more parochial will strengthen constraints on behavior through reputational effects for the ingroup and weaken these constraints for the outgroup. We will come back to this point when discussing the observer.

Power asymmetries. Power asymmetries are likely to generate strong effects of salient hedonic goals (sexual abuse, fear) and to change the working of the normative goal. When people in one group differ much in terms of power, it is likely that the more powerful will care less about reputational effects than the less powerful. This is not just because power can neutralize the negative consequences of reputational effects, but also because power can be used to either prevent reputational effects from occurring, or to exploit them. With relational asymmetries, the powerful are also more likely to be less concerned about using unethical means (Brass, Butterfield and Skaggs 1998). Importantly, the powerful can create fear in the less powerful about spreading certain information about them, purposefully building up a reputation for being fierce and unforgiving. This reigning by fear is made less effective though by the ability of the "weak" to use reputational effects as a weapon, by spreading rumors that the powerful would rather not have spread. In this case, people don't spread first-hand information but rumors they say they heard from others. Such hear-say has no identifiable source and no truth claim, and thus it is less likely to be punishable (Giardini, 2012). This form of gossip can be a strong weapon of the

weak against the powerful. The downside of this weapon is that it lacks some bite by missing identifiable sources and truth claims.

The powerful can, in turn, neutralize much of the damaging effect of gossip by interpreting behavior in such a way that the working of the normative goal is undermined. They can create moral ambiguity by reacting lukewarm or even with praise to the blatant infractions of norms and standards they wish to undermine or the applicability of which they wish to control. For example, president Trump reacted to blatant white racial violence in Charlottesville in 2017 by saying "I think the blame is on both sides [. . .] You had a group on one side that was bad. You had a group on the other side that was also very violent. Nobody wants to say that. I'll say it right now." (*New York Times*, Aug. 11th, 2017).

The powerful can also influence the interpretation of behavior in ways that allow the direct manipulation of information, including gossip. One way has been described by Orwell in his novel 1984 as the imposition of "doublethink" and "newspeak". He defines the former as "the act of holding, simultaneously, two opposite, individually exclusive ideas or opinions and believes" and the latter as "an official or semiofficial style of writing or saying one thing in the guise of its opposite". There are many present-day examples of these methods, such as the introduction of the term "alternative facts" by the advisor to the White House.

Another (but related) way to influence the interpretation of behavior and create moral ambiguity is to redefine the behavior as not falling under injunctive (that is "obligatory") norms but under descriptive norms (what people do) (Lindenberg et al 2020). This is achieved by using one's power to "persuade" what would otherwise be victims of transgression of injunctive norms that they are in unfamiliar terrain and thus have to learn how things are done around here (descriptive norms). Descriptive norms can undermine injunctive norms or at least their applicability. For example, a powerful film producer can persuade actresses looking for a role in a movie that it is normal, for getting a chance of being considered for a part, to be willing to play sexual games with the producer. Ironically, hearing that everybody does it does not cumulatively add to the bad reputation of the producer but makes it more likely that women submit to him. Thus, the power to refocus the interpretation of behavior from being subject to injunctive norms to being subject to manipulated descriptive norms also diminishes whatever reputational effects might have achieved. Evaluative judgments about descriptive norms refer to personal preference, similar to what people might think about chewing gum ("I hate it, but seemingly everybody does it") rather than to moral condemnation. Thus, gossip about personal preferences will not have much impact on reputations.[8]

8 Game-theoretic approaches that take norms as equilibrium behavior (that is what most people do) treat all norms as though they were descriptive norms and thus cannot even describe such shifts in evaluative judgments from moral condemnation to personal preference.

6.6 The observer

Reputational effects, especially for actors with a salient gain goal, can restrain behavior because there are presumably observers who might react negatively or pass on to others what they have experienced. Observers can be targets of actions that trigger an evaluative response or they can be observers of such actions happening to others. For both, it is not trivial whether or not some action or information is interpreted positively or negatively. There may be sheer flaws in the system providing information. For example, "likes" and positive evaluations for one's services can be freely bought on the internet. Observers of feedback on social media may or may not be able to distinguish fake from genuine reviews. Up to now, we know too little about the ability to pinpoint fake reviews. Government agencies, concerned about the bad quality of the online reputations systems, have taken action against fake reviews. For example, the Competition & Market Authority of the British Government issued an open letter to marketing departments and agencies and their clients warning about fake reviews, and reporting that it has taken enforcement action against widespread fake online reviews (CMA 2016). But besides such faults in the feedback systems, there are important psychological aspects that influence how an observer interprets actions or information.

6.7 Special effects for the observer: Ingroup/outgroup and power-asymmetry effects

Ingroup/outgroup effects. Reputation effects necessitate the interpretation of others' behavior as transgression of norms. However, ingroup/outgroup dynamics bias the interpretation of behavior as transgression, because cognition is affected by motivation (Brewer 1979, Balcetis and Dunning 2006). Stronger parochialism will make it more likely that actions by an outgroup member are interpreted as transgressions of the ingroup norms and standards. There is little room for granting outgroup members the possibility of having made mistakes or of misunderstandings, a privilege that is often granted to ingroup members. Even if one is not a target, observing of members of one's ingroup being badly treated by members of an outgroup is likely to trigger moral outrage (Veldhuis et al. 2014). Conversely, the stronger parochialism, the less likely an observer will interpret a norm transgression of an ingroup members as a transgression (Hughes et al. 2017, Everett et al. 2015). More likely it will be interpreted as an error, or an action caused by an outsider. For example, seemingly, people who see Donald Trump as one of their ingroup members do not believe the accusations about his sexual transgressions ("fake news"), or see it as a forgivable mistake ("everybody makes mistakes"). In this way transgressions fail to contribute to a bad reputation among the ingroup members.

If, however, an ingroup member must admit that the observed behavior by a fellow ingroup member was a transgression, a likely reaction is that the "culprit" is not, should not be, or never was a true member of the ingroup. The reputation effect is then socially fatal. For example, in the wake of the Harvey Weinstein scandal and the "me too" tweets, famous actors who were then also accused of sexual misconduct were removed from television series, as if watching them would be unbearable. This taps into the purity dimension of morality (see Graham et al. 2011) and shows that reputation effects often have much to do with increased salience of the hedonic goal and little to do with a rational calculation of risk in dealing with a particular person. It also reinforces that, contrary to what Granovetter and Raub assume, reputation is not only about finding out whether somebody is or is not a trustworthy type. There is also fear of contagion, fear that "badness" rubs off, that one may be corrupted by being near or even just seeing such a person. In short, to the degree that reputation effects depend on observers, parochialism dampens these effects concerning the ingroup members, and distorts these effects concerning outgroup members. In both cases, parochialism reduces the information value carried by reputations or the lack of them.

Power asymmetry effects. Like ingroup/outgroup dynamics, power asymmetry affects the perception of the transgressions of others. Being powerful goes with a strong tendency to focus on achieving one's goals (Guinote 2007). If the observer is powerful, noticing others' transgressions depends on their relation to the observer's goal-pursuit. If the transgression is deemed relevant, it is most likely observed and remembered by the powerful, leading potentially to reputation effects. If the transgressor is perceived as being low in power, the chance is small that the transgression touches the goal pursuit of the powerful person. Thus, as long as persons low in power do not cross the goal pursuit of a person high in power, their transgressions are likely not even noticed by the powerful, highly limiting reputation effects.

It is different if the observer is low in power. Then, observing transgressions of a powerful person may be keenly noticed but is subject to reappraisal. "Did he just insult me? Well, probably not. I think he has a bad day". Individuals with low power are more perceptive of context effects and threats emanating from contexts than individuals with high power (Kraus et al. 2012). This makes them more likely to interpret transgressions of more powerful others as being more harmless that they were (that is they reappraise, see Hittner, Rim and Haase 2018). In this way, people are on the safe side, because what they saw does not call for a reaction. But then it also does not lead to a potential reputation affect.

In the business world, there is an additional way in which the powerful may escape being subject to possible negative reputation effects. Big companies can afford investing considerable amounts of resources in branding and high visibility, gaining consumer's trust with little help from reputational effects. By contrast, small companies must rely on mechanisms of reputation building (such as online

peer review systems) to gain consumer's trust. The functioning of such reputation systems thus has a strong impact on the power asymmetry of companies. For example, Newberry & Zhou (2019) show that if the reputation system for small companies would not be available, there would be a large shift in demand for goods from the small to the big companies. This makes understanding the conditions under which reputation systems work all the more important.

6.8 The transmitter

As Burt (2008) observed: "reputations emerge not form what we do, but from people talking about what we do [. . .] What circulates depends on the interest of people doing the circulation." The transmitter deserves special attention with regard to reputation effects.

An observer may or may not be also a transmitter, that is someone who passes on what he or she has experienced. Transmitters may also be people who pass on not what they have experienced but what they have heard or say they have heard. The important question is why people would become transmitters. Obviously, they may have a strategic goal in mind, to help or hurt somebody by doing so. For example, a person with a salient normative goal may want to help the authorities by passing on information on a criminal act he has seen somebody commit (Feinberg et al. 2012). A person with a salient gain goal may want to hurt the competition by passing on negative information about the competitor. A person with a salient hedonic goal may want to take revenge on somebody by disclosing negative confidential information.

Targets, especially victims, are likely to react emotionally to infractions and thus have an increased salience of the hedonic goal, which means that they have less strategic restraint, and that they are likely to choose actions that promise to improve the way they feel (such as retaliation, aggression in case of victims, and the pleasure of being nasty to people they don't like) (Barclay, Skarlicki, and Pugh, 2005). Because of their salient hedonic goal, they are potentially very important transmitters. Taking revenge makes people feel better (Chester and De Wall 2017), and telling the world is often part of taking revenge. It is these kinds of hedonic responses to having become a victim that are dangerous for actors. However, hedonic reactions can be curbed by hedonic means. Thus, instilling fear or shame in victims is a powerful way to keep victims from becoming transmitters. We mentioned already the example of buyers on eBay who fear negative counter punishment (Li 2010). Harassment in the workplace (prominently including sexual harassment) is also a good example of this. A recent study by the U.S. Equal Employment Opportunity Commission (Feldblum and Lipnic 2016), reported that about 75% of employees who experienced harassment never even talked to a supervisor, manager, or union representative about the harassing

conduct. "Employees who experience harassment fail to report the harassing behavior or to file a complaint because they fear disbelief of their claim, inaction on their claim, blame, or social or professional retaliation." (Feldblum and Lipnic 2016: v).

Gossip. Even though fear will have a smaller impact on gossipers than on victims because hear-say has no identifiable source and no truth claim (Giardini and Conte 2012), gossipers are also likely to have a salient hedonic goal and are thus sensitive to cues that instill fear.

People like to make conversation, small talk, and have something interesting to say or hear in the process. Importantly, when people are in a situation that invites behavior that is pleasant or fun, such as gossiping, their hedonic goal becomes more salient, making it also a pleasant experience to feel virtuous by judging others without having to sacrifice for being virtuous oneself (Lindenberg et al. 2018). In such a situation, people tend to feel more virtuous than others (Epley and Dunning 2000). In addition, it is likely that people gladly take any opportunity to engage in this behavior because gossip is able to satisfy all five fundamental needs (Nieboer et al. 2005), often at the same time: a need for stimulation; a need for comfort, a need for affection (belonging); a need for behavioral confirmation; and a need for status. For example, by telling someone confidentially about how bad somebody behaved, the gossiper provides stimulation for the other, some status for himself for being in the know, some behavioral confirmation for himself and the other as they exchange disapproval of a disgusting act "they would never do"; some bonding by sharing information others don't have, and some comfort by mutually feeling the other's support.

Gossip does not have to be about information that affects somebody's reputation. For example, it could be about the incredible bad luck for somebody else, rather than bad deeds. This is what De Backer et al. (2019) call "strategy learning gossip" (see also Baumeister, Zhang and Vohs 2004), as opposed to "reputation gossip" which is about people known to the gossiper, directly or indirectly. Still, gossip is potentially a major vehicle by which reputationally relevant information and rumors circulate. Gossip flows. The motivational force of gossip that is based on the satisfaction of fundamental needs is quite independent of the wish to help or hurt specific others, and it is largest when the gossip is about negative behavior of others, because that serves best the needs for stimulation, behavioral confirmation, and status (Baumeister et al. 2001). Yet, this high potential of gossip is often not realized, because it is likely to be only about outsiders and kept local by the boundaries of an inner circle.

6.9 Special effects for the transmitter: Ingroup/outgroup effects

Even though gossip is potentially a most important vehicle for reputation effects, it is likely to remain very local and thus without much bite. For example, it has been

shown that within the ingroup in organizations, positive and negative gossip is shared, whereas with outgroup members, only positive gossip is shared (Grosser, Lopez-Kidwell, and Labianca, 2010). This restriction of gossip has been generally overlooked by those who praise the importance of gossip for reputations effects and general human cooperation (for a review on gossip, reputation and cooperation see Giardini and Wittek, 2019). Why is gossip likely to be highly restricted in its reach?

True to the fission/fusion dynamics of human group formation, there are various ingroups to which one may flexibly belong, ranging from categorical groups ("we women") to countries, to ever smaller social units, and finally to "inner circles" of people who confide in each other. These inner circles play an important role with regard to negative gossip. To badmouth somebody can be dangerous, if one is not sure that the other shares one's evaluations. This is danger that may not come from fear of a powerful third party but from being rebuffed, from losing behavioral confirmation and affection (Cole and Scriverner 2013). This is one reason why negative gossip, especially that about people not belonging to the one's own circle, circulates mainly in circles of close ties (Travis et al. 2010; McAndrew and Milenkovic 2002; Milliken, Morrison and Hewlin 2003). Another reason for keeping negative gossip about outgroup members to the ingroup is the bonding function of gossip (Dunbar, 1996; Foster 2004). By exchanging negative gossip about an outsider, the bond between the gossipers is strengthened, mutually reinforcing the membership in the same inner circle. By exchanging negative gossip with outsiders, this function is not fulfilled.

6.10 Power asymmetry effects: Broken by a special transmitter, the publicizer

This close circle restriction of gossip is broken when a person that is negatively gossiped about is already widely seen in a negative light. Then negative gossip is no longer dangerous and flows freely. It is then that it may lead to socially fatal reputation effects. This is likely to happen in cases where power asymmetries strengthened the ingroup/outgroup dynamics of gossip and where a publicizer was able to spread negative news about the powerful. For example, in case of Harvey Weinstein, all hell broke loose, once he was branded negatively and it was safe to negatively gossip anywhere about him, creating a cascade of negative news in the "me too" movement. In this case, the ingroup of the inner circle expanded to the world-wide good guys against the bad guys. Another example is Gorbachev's televised speech on November 3rd 1987, in which he attacked Stalin, saying that continued neglect of Stalin's crimes was unacceptable. Stalin was guilty of "enormous and unforgivable" crimes. "Many thousands of people inside and outside the party were subjected to wholesale repressive measures." (New York Times Nov. 3rd, 1987). Now that a publicizer made Stalin stand

publicly in a negative light, masses of people began to speak out in public about the horrible acts they had experienced under Stalin.

Publicizers have a special role. They are more than normal whistleblowers. They may have to be powerful enough to withstand the counter pressure of those who suppressed reputation effects in the first place. Publicizers expand the reach of reputation effects by breaking the inner circle restriction of gossip. But just how far they are able to expand the reach depends on the kind of publicizer and the news value of the revelation. Powerful politicians (such as Gorbachev) and powerful media (such as the *New York Times*) are probably the publicizers who can achieve the largest reach and also overcome people's reluctance to believe bad things about a person previously in good standing. The power to control the media is thus a mighty weapon against the socially fatal effects of a bad reputation, as every dictator knows so well. But because publicizers mostly focus on people with actual or potential fame and most likely on extreme forms of deviant behavior (such as pedophilia, sexual abuse of women, downright corruption), for "ordinary" people, reputation effects remain often confined to what circulates in inner circles about outsiders, thereby being only loosely related to what people actually did (Anderson and Shirako 2008).

6.11 Conclusion

The importance of reputation effects for social order makes it mandatory for research on this terrain to use all the tools at the disposal of social and behavioral sciences to come to understand their dynamics. This means paying close attention to normative embeddedness (and normative heterogeneity), to structural features beyond networks (ingroup/outgroup dynamics and power differences), and to the dynamics of overarching goals that govern virtually all conditions under which reputation effects are likely to occur. Granovetter (1985) insisted that for a full understanding of reputation effects, we have to consider dyadic and structural embeddedness and all we need as a behavioral theory is the assumption of rational choice. This, we argued, was a grave error. Granovetter thought that structural embeddedness together with rational choice would finally sideline the atomistic explanations of social order on the basis of internalized norms and would also sideline "psychological revisionism" that behavioral economics had increasingly engaged in. He did not think of normative embeddedness, nor did he think of structures beyond social networks, and he completely neglected the importance of overarching goals. Werner Raub is one of the major figures who embraced and advanced Granovetter's embeddedness approach. However, he repeated Granovetter's error, thereby limiting what can be learned about reputation effects from his extended research program.

We show that reputation effects rest on normative embeddedness and that this embeddedness does not just consist of shared norms but also of the dynamics of

overarching goals that affect how and when norms guide behavior. A salient overarching goal governs what we pay attention to and what we ignore, what alternative we consider, what information we are sensitive to, what we like and dislike etc. In short, overarching goals govern the cognitive and motivational processes that are most important for behavior. Attention to these overarching goals is important with regard to virtually every aspect of reputation effects, especially for answering the question how it is possible that people make negative judgments about other people's behavioral tendencies on the basis of one or a few observations.

The "normative" goal to act appropriately, follow norms, be oriented to the collective, is an overarching goal the salience of which shifts with social and institutional circumstances, including structural embeddedness. Even if people have internalized norms, they will not act on the basis of these norms unless the normative goal is salient at the moment, and this saliency, in turn, depends on the structural embedding, so that normative and structural embeddedness interact. Next to the normative overarching goal, there are two competing overarching goals: one focused on resources (gain), and one focused on feeling good (hedonic). If all people permanently had a salient gain goal (that is be rational egoists), then being observed acting egoistically by lying, cheating, not helping etc. would have no news value, would not contribute to one's reputation. A bad reputation is acquired by a presumed tendency not to have a salient normative goal when this is socially expected. By contrast, one does not acquire a bad reputation by having a salient gain goal in a price battle with a competitor. Reputation is based on an incongruence between observed and expected salience of an overarching goal.

Normative embeddedness is not just sharing norms and standards, but also being subject to the conditions that affect the salience of the normative goal, and this also affects the actor, the observer of other people's actions, as well as the transmitter of information or rumors about these actions. Especially important are heterogeneity with regard to norms, ingroup/outgroup dynamics, and power asymmetries. All three affect the salience of overarching goals and thereby the likelihood that people's behavior is constrained by anticipated reputation effects, the likelihood that negative behavior will be talked about, and the likelihood that what is said about people is actually informative or highly distorted. In sum, the study of reputation effects necessitate that we grant center stage to normative embeddedness and its interaction with structural embeddedness. Granovetter is exemplary in much of his work, but his error should be seen as a strong obstacle for the analysis of reputation effects.

References

Anderson, Cameron, and Aiwa Shirako. "Are Individuals' Reputations Related to Their History of Behavior?" *Journal of Personality and Social Psychology* 94 (2) (2008): 320.

Aureli, Filippo, Colleen M. Schaffner, Christophe Boesch, Simon K. Bearder, Josep Call, Colin A. Chapman, Richard Connor, Anthony Di Fiore, Robin IM Dunbar, and S. Peter Henzi. "Fission-Fusion Dynamics: New Research Frameworks." *Current Anthropology* 49 (4)(2008): 627–654.

Balcetis, Emily, and David Dunning. "See What You Want to See: Motivational Influences on Visual Perception." *Journal of Personality and Social Psychology* 91 (4) (2006): 612.

Balmer, John MT, and Edmund R. Gray. "Corporate Identity and Corporate Communications: Creating a Competitive Advantage." *Corporate Communications: An International Journal* 4 (4) (1999): 171–177.

Barclay, Laurie J., Daniel P. Skarlicki, and S. Douglas Pugh. "Exploring the Role of Emotions in Injustice Perceptions and Retaliation." *Journal of Applied Psychology* 90 (4) (2005): 629.

Baumeister, Roy F., Ellen Bratslavsky, Catrin Finkenauer, and Kathleen D. Vohs. "Bad Is Stronger than Good." *Review of General Psychology* 5 (4) (2001): 323–370.

Baumeister, Roy F., Liqing Zhang, and Kathleen D. Vohs. "Gossip as Cultural Learning." *Review of General Psychology* 8 (2) (2004): 111–121.

Birkinshaw, Julian, Nicolai Foss, and Siegwart Lindenberg. "Purpose with Profits: How to Make Your pro-Social Goals Pay." *MIT Sloan Management Review* 55 (3) (2014): 49–56.

Bolton, Gary, Ben Greiner, and Axel Ockenfels. "Engineering Trust: Reciprocity in the Production of Reputation Information." *Management Science* 59 (2) (2013): 265–285.

Bolton, Gary, Ben Greiner, and Axel Ockenfels. "Dispute Resolution or Escalation? The Strategic Gaming of Feedback Withdrawal Options in Online Markets." *Management Science* 64 (9) (2017): 4009–4031.

Brass, Daniel J., Kenneth D. Butterfield, and Bruce C. Skaggs. 1998. "Relationships and Unethical Behavior: A Social Network Perspective." *Academy of Management Review* 23 (1): 14–31.

Brewer, Marilynn B. "In-Group Bias in the Minimal Intergroup Situation: A Cognitive-Motivational Analysis." *Psychological Bulletin* 86 (2) (1979): 307.

Bryan, Christopher J., Gabrielle S. Adams and Benoit Monin. "When Cheating Would Make You a Cheater: Implicating the Self Prevents Unethical Behavior." *Journal of Experimental Psychology: General* 142, No. 4 (2013): 1001–1005

Burt, Ronald S., and Pietro Panzarasa. "Gossip and Reputation." Pp. 27–42 in *Management et Réseaux Sociaux: Ressource Pour l'action Ou Outil de Gestion*, edited by Marc Lecoutre and Lievre Pascal. London: Hermes-Lavoisier, 2008.

Buskens, V., and W. Raub. "Rational Choice Research on Social Dilemmas: Embeddedness Effects on Trust" Pp. 113–150 in *The Handbook of Rational Choice Social Research*, edited by Rafael Wittek, Tom A. B. Snijders and Victor Nee. Stanford, California: Stanford University Press, 2013.

Callahan, David. *The Cheating Culture: Why More Americans Are Doing Wrong to Get Ahead*. New York: Houghton Mifflin Harcourt, 2004.

Chester, David S., and C. Nathan DeWall. "Combating the Sting of Rejection with the Pleasure of Revenge: A New Look at How Emotion Shapes Aggression." *Journal of Personality and Social Psychology* 112 (3) (2017): 413.

Cohn, Alain, Ernst Fehr, and Michel André Maréchal. "Business Culture and Dishonesty in the Banking Industry." *Nature* 516 (7529) (2014): 86.

Cole, Jennifer M., and Hannah Scrivener. "Short Term Effects of Gossip Behavior on Self-Esteem." *Current Psychology* 32 (3) (2013): 252–260.

Coleman, James S. "Collective Decisions." *Sociological Inquiry* 34 (2) (1964): 166–181.

Competition & Markets Authority (CMA). "An open letter to marketing departments, marketing agencies and their clients." 11. 8.2016,Retrieved 6.7.2018 from https://www.icpen.org/initiatives#industry-guidance

Corten, Rense, Stephanie Rosenkranz, Vincent Buskens, and Karen S. Cook. "Reputation Effects in Social Networks Do Not Promote Cooperation: An Experimental Test of the Raub & Weesie Model." *PloS One* 11 (7) (2016): e0155703.

Dana, Jason, Roberto A. Weber, and Jason Xi Kuang. "Exploiting Moral Wiggle Room: Experiments Demonstrating an Illusory Preference for Fairness." *Economic Theory* 33 (1) (2007): 67–80.

Davis, Adam, Tracy Vaillancourt, Steven Arnocky, and Robert Doyel. "Women's Gossip as an Intrasexual Competition Strategy." Pp. 303- in *The Oxford Handbook of Gossip and Reputation*, edited by Francesca Giardini and Rafael Wittek. Oxford: Oxford University Press, 2019.

De Backer, Charlotte JS, Hilde Van den Bulck, Maryanne L. Fisher, and Gaëlle Ouvrein. "Gossip and Reputation in the Media." Pp. 325- in *The Oxford Handbook of Gossip and Reputation*, edited by Francesca Giardini and Rafael Wittek. Oxford: Oxford University Press, 2019.

De Dreu, Carsten KW, Jörg Gross, Zsombor Méder, Michael Giffin, Eliska Prochazkova, Jonathan Krikeb, and Simon Columbus. 2016. "In-Group Defense, out-Group Aggression, and Coordination Failures in Intergroup Conflict." *Proceedings of the National Academy of Sciences* 113 (38): 10524–10529.

Diekmann, Andreas, Ben Jann, Wojtek Przepiorka, and Stefan Wehrli. "Reputation Formation and the Evolution of Cooperation in Anonymous Online Markets." *American Sociological Review* 79 (1) (2014): 65–85.

Dunbar, Robin. *Grooming, Gossip and the Evolution of Language* (1996). London: Faber and Faber.

Dunning, David. "Normative Goals and the Regulation of Social Behavior: The Case of Respect." *Motivation and Emotion* 41 (3)(2017): 285–293.

Durkheim, Emile. *The Division of Labor in Society*. New York: Free Press, 1964 (originally 1893 in French).

Dyer, Jeffrey H., and Harbir Singh. "The Relational View: Cooperative Strategy and Sources of Interorganizational Competitive Advantage." *Academy of Management Review* 23 (4) (1998): 660–679.

Engelmann, Jan M., Esther Herrmann, Michael Tomasello. "Five-year olds, but not chimpanzees, attempt to manage their reputations". *PLOS ONE*, *7*(10), (2012) https://doi.org/10.1371/journal.pone.0048433

Epley, Nicholas, and David Dunning. "Feeling 'holier than thou": Are self-serving assessments produced by errors in self- or social prediction?" *Journal of Personality and Social Psychology*, 79(6), (2000): 861–875

Evans, Anthony M., and Philippe P. F. M. van de Calseyde. "The Reputational Consequences of Generalized Trust." *Personality and Social Psychology Bulletin* 44, no. 4 (April 2018): 492–507. doi:10.1177/0146167217742886.

Everett, Jim A.C., Nadira S. Faber and Molly J. Crockett. "The influence of social preferences and reputational concerns on intergroup prosocial behaviour in gains and losses contexts." *Royal Society Open Science* 2 (2015): 150546. http://dx.doi.org/10.1098/rsos.150546

Feinberg, Matthew Aaron, Robb Willer, Jennifer E. Stellarand Dacher Keltner. "The virtues of gossip: reputational information sharing as prosocial behavior." *Journal of Personality and Social Psychology* 102, 5 (2012): 1015–30

Fehr, Ernst, Martin Brown, and Christian Zehnder. "On Reputation: A Microfoundation of Contract Enforcement and Price Rigidity", *The Economic Journal* 119, Issue 536 (March 2009): 333–353, https://doi.org/10.1111/j.1468-0297.2008.02240.x

Feldblum, Chai R., and Victoria A. Lipnic. "Report of the co-chairs of the EEOC Select Task Force on the Study of Harassment in the Workplace". Washington, D.C.: U.S. Equal Employment Opportunity Commission (2016).

Foster, Eric K. "Research on gossip: Taxonomy, methods, and future directions". *Review of General Psychology*, *8*, (2004): 78–99.

Gallup *Honesty/Ethics in Professions*. https://news.gallup.com/poll/1654/honesty-ethics-professions.aspx (accessed May 25, 2018), 2014.

Giardini, Francesca. "Deterrence and transmission as mechanisms ensuring reliability of gossip". *Cognitive Processing* 13(2), (2012): 465–475.

Giardini, Francesca, and Rafael Wittek. "Gossip, Reputation, and Sustainable Cooperation: Sociological Foundations." In *The Oxford Handbook of Gossip and Reputation*, edited by Francesca Giardini, and Rafael Wittek, 23–47. New York: Oxford University Press, 2019.

Giardini, Francesca, and Rosaria Conte. "Gossip for social control in natural and artificial societies", *Simulation: Transactions of the Society for Modeling and Simulation International* 88(1), (2012): 18–32

Gintis, Herbert, and Ernst Fehr. "The social structure of cooperation and punishment." *Behavioral and Brain Sciences* 35, no. 1 (2012): 28–29.

Gintis, Herbert, Eric Alden Smith, and Samuel Bowles. "Costly signaling and cooperation." *Journal of Theoretical Biology* 213, no. 1 (2001): 103–119.

Guennif, Samira and Valerie Revest. "Social structure and reputation: the NASDAQ case study. *Socio-Economic Review 3* (2005): 417–436

Graham, Jesse, Brian A. Nosek, Jonathan Haidt, Ravi Iyer, Spassena Koleva, and Peter H. Ditto. "Mapping the moral domain." *Journal of Personality and Social Psychology* 101, no. 2 (2011): 366.

Granovetter, Mark. "Economic action and social structure: The problem of embeddedness." *American journal of sociology* 91, no. 3 (1985): 481–510.

Granovetter, Mark. "Problems of explanation in economic sociology." In *Networks and Organizations: Structure, Form, and Action*, edited by Nohria, Nitin, and Richard Eccles, 25–56. Boston, MA: Harvard Business School Press, 1992.

Gray, Edmund R., and John MT Balmer. "Managing corporate image and corporate reputation." *Long Range Planning* 31, no. 5 (1998): 695–702.

Grosser, Travis J., Virginie Lopez-Kidwell, and Giuseppe Labianca. 2010. "A Social Network Analysis of Positive and Negative Gossip in Organizational Life." *Group & Organization Management* 35 (2): 177–212.

Guinote, Ana. "Power and goal pursuit." *Personality and Social Psychology Bulletin* 33, no. 8 (2007): 1076–1087.

Hilbe, Christian, Moshe Hoffman, and Martin Nowak. "Cooperate without looking in a non-repeated game." *Games* 6, no. 4 (2015): 458–472.

Hittner, Emily F., Katie L. Rim, and Claudia M. Haase. "Socioeconomic Status as a Moderator of the Link Between Reappraisal and Anxiety: Laboratory-Based and Longitudinal Evidence." *Emotion* (2018). Online First Publication, December 17, 2018. http://dx.doi.org/10.1037/emo0000539

Hoffman, Moshe, Erez Yoeli, and Carlos David Navarrete. **"**Game Theory and Morality." In *The Evolution of Morality*, edited by Todd K. Shackelfordand Roger D. Hansen, 289–316. New York: Springer, 2016.

Homans, George C. "Bringing Men Back in. *American Sociological Review* 29 (1964): 809–818.

Hughes, Brent T., Jamil Zaki, and Nalini Ambady. "Motivation alters impression formation and related neural systems." *Social Cognitive and Affective Neuroscience* 12(1) (2017): 49–60

Jordan, Jillian J., Moshe Hoffman, Martin A. Nowak, and David G. Rand. "Uncalculating cooperation is used to signal trustworthiness." *PNAS* 113(31) (2016):8658–63. doi: 10.1073/pnas.160128011

Keizer, Kees, Siegwart Lindenberg, and Linda Steg. "The spreading of disorder." *Science* 322, no. 5908 (2008): 1681–1685.

Keizer, Kees, Siegwart Lindenberg, and Linda Steg. "The importance of demonstratively restoring order." *PloS ONE* 8, no. 6 (2013): e65137.

Kraus, Michael W., Paul K. Piff, Rodolfo Mendoza-Denton, Michelle L. Rheinschmidt, and Dacher Keltner. "Social class, solipsism, and contextualism: how the rich are different from the poor." *Psychological Review* 119, no. 3 (2012): 546.

Lawrence, Thomas B. "Examining resources in an occupational community: Reputation in Canadian forensic accounting." *Human Relations* 51, no. 9 (1998): 1103–1131.

Li, Lingfang. "What is the Cost of Venting? Evidence from eBay." *Economics Letters* 108, no. 2 (2010): 215–218.

Lindenberg, Siegwart. "Zur Kritik an Dürkheims Programm für die Soziologie." *Zeitschrift für Soziologie* 12, no. 2 (1983): 139–151.

Lindenberg, Siegwart. "Complex Constraint Modeling (CCM): A Bridge Between Rational Choice and Structuralism: Comment." *Journal of Institutional and Theoretical Economics (JITE)/Zeitschrift für die gesamte Staatswissenschaft* 151, no. 1 (1995): 80–88.

Lindenberg, Siegwart. "It takes both trust and lack of mistrust: The workings of cooperation and relational signaling in contractual relationships." *Journal of Management and Governance* 4, no. 1–2 (2000): 11–33.

Lindenberg, Siegwart. "Social norms: What happens when they become more abstract?." In *Rational Choice: Theoretische Analysen und empirische Resultate*, pp. 63–81. Wiesbaden: VS Verlag für Sozialwissenschaften, 2008.

Lindenberg, Siegwart. "How cues in the environment affect normative behavior." In *Environmental Psychology: An Introduction*, edited by Linda Steg, A.E. van den Berg, & J.I.M. de Groot, 119–128. New York: Wiley, 2012.

Lindenberg, Siegwart. "The Dependence of Human Cognitive and Motivational Processes on Institutional Systems." Pp.85–106 in Ben Jann & Wojtek Przepiorka (eds.) *Social Dilemmas, Institutions and the Evolution of Cooperation,* Berlin: De Gruyter Oldenbourg, 2017.

Lindenberg, Siegwart, and Linda Steg. "Normative, Gain and Hedonic Goal Frames Guiding Environmental Behavior." *Journal of Social Issues* 63 (1) (2007): 117–137.

Lindenberg, Siegwart, and Linda Steg. 2013. "Goal-Framing Theory and Norm-Guided Environmental Behavior." *Encouraging Sustainable Behavior*, New York: Psychology Press.

Lindenberg, Siegwart, Linda Steg, Marko Milovanovic, and Anita Schipper. "Moral Hypocrisy and the Hedonic Shift: A Goal-Framing Approach." *Rationality and Society* 30 (4) (2018): 393–419.

Lindenberg, Siegwart, Frederique Six, and Kees Keizer. "Social contagion and goal framing: The sustainability of rule compliance." Chapter 29 in D. Daniel Sokol and Benjamin van Rooij (eds.). *Cambridge Handbook of Compliance*. Cambridge: Cambridge University Press, 2020.

Manesi, Zoi, Paul AM Van Lange, and Thomas V. Pollet. "Eyes Wide Open: Only Eyes That Pay Attention Promote Prosocial Behavior." *Evolutionary Psychology* 14 (2) (2016): 1–15, 1474704916640780.

Mazar, Nina, On Amir, and Dan Ariely. "The Dishonesty of Honest People: A Theory of Self-Concept Maintenance." *Journal of Marketing Research* 45 (6) (2008): 633–644.

McAndrew, Francis T., and Megan A. Milenkovic. "Of Tabloids and Family Secrets: The Evolutionary Psychology of Gossip 1." *Journal of Applied Social Psychology* 32 (5) (2002): 1064–1082.

Milinski, Manfred. "Reputation, a Universal Currency for Human Social Interactions." *Philosophical Transactions of the Royal Society B: Biological Sciences* 371 (1687) (2016): 20150100.

Milliken, Frances J., Elizabeth W. Morrison, and Patricia F. Hewlin. "An Exploratory Study of Employee Silence: Issues That Employees Don't Communicate Upward and Why." *Journal of Management Studies* 40 (6) 2003: 1453–1476.

Newberry, Peter and Xiaolu Zhou. "Heterogeneous effects of online reputation for local and national retailers." *International Economic Review* 60 (4) (2019):1 564–1587

Nieboer, Anna, Siegwart Lindenberg, Anne Boomsma, and Alinda C. Van Bruggen. "Dimensions of Well-Being and Their Measurement: The SPF-IL Scale." *Social Indicators Research* 73 (3) (2005): 313–353.

O'Mara, Erin M., Lydia E. Jackson, C. Daniel Batson, and Lowell Gaertner. "Will Moral Outrage Stand up?: Distinguishing among Emotional Reactions to a Moral Violation." *European Journal of Social Psychology* 41 (2) (2011): 173–179.

Preece, J., D. Maloney-Krichmar, and C. Abras. "History of Online Communities In Karen Christensen & David Levinson (Eds.), Encyclopedia of Community: From Village to Virtual World." *Thousand Oaks: Sage Publications Anorexic Patients' Competence. Archives of Psychiatry & Psychotherapy* 19 (2003): 39–43.

Pulford, Briony D., Eva M. Krockow, Andrew M. Colman, and Catherine L. Lawrence. "Social Value Induction and Cooperation in the Centipede Game." *PLOS ONE* 2016 | DOI:10.1371/journal.pone.0152352

Raub, Werner, and Vincent Buskens. "Theory and Empirical Research in Analytical Sociology: The Case of Cooperation in Problematic Social Situations." *Analyse & Kritik* 30 (2) (2008): 689–722.

Raub, Werner, and Jeroen Weesie. "Reputation and Efficiency in Social Interactions: An Example of Network Effects." *American Journal of Sociology* 96 (3) (1990): 626–654.

Schilke and Karen Cook. "Sources of alliance partner trustworthiness: integrating calculative and relational perspectives." *Strategic Management Journal* 36 (2015): 276–297

Shah, James. "Automatic for the People: How Representations of Significant Others Implicitly Affect Goal Pursuit." *Journal of Personality and Social Psychology* 84 (4) (2003a): 661.

Shah, James. "The Motivational Looking Glass: How Significant Others Implicitly Affect Goal Appraisals." *Journal of Personality and Social Psychology* 85 (3) (2003b): 424.

Simpson, Brent, and Robb Willer. "Altruism and Indirect Reciprocity: The Interaction of Person and Situation in Prosocial Behavior." *Social Psychology Quarterly* 71 (1) (2008): 37–52.

Spiekermann, Kai, and Arne Weiss. "Objective and Subjective Compliance: A Norm-Based Explanation of 'Moral Wiggle Room.'" *Games and Economic Behavior* 96 (2016): 170–183.

Traister, Rebecca. "Why the Harvey Weinstein Sexual-Harassment Allegations Didn't Come Out Until Now." *The Cut*, Oct. 5th, 2017 Retrieved from the internet 8.7.2018 https://www.thecut.com/2017/10/why-the-weinstein-sexual-harassment-allegations-came-out-now.html

Uzzi, Brian. "Social Structure and Competition in Interfirm Networks: The Paradox of Embeddedness". *Administrative Science Quarterly* 42(1) (1997): 35–67.

Veldhuis, Tinka M., Ernestine H. Gordijn, Rene Veenstra, and Siegwart Lindenberg. "Vicarious Group-Based Rejection: Creating a Potentially Dangerous Mix of Humiliation, Powerlessness, and Anger." *PloS One* 9 (4) (2014): e95421.

Wedekind, Klaus and Manfred Milinksi. "Cooperation through image scoring in humans." *Science* 288 (2000): 850–852.

Weigelt, Keuth & Colin Camerer. "Reputation and Corporate Strategy." *Strategic Management Journal*, 9(5) (1988): 443–454.

Wetzer, Inge M., Marcel Zeelenberg, and Rik Pieters. "'Never Eat in That Restaurant, I Did!': Exploring Why People Engage in Negative Word-of-Mouth Communication." *Psychology & Marketing* 24 (8) (2007): 661–680.

Wildschut, Tim, and Chester A. Insko. "A Paradox of Individual and Group Morality: Social Psychology as Empirical Philosophy." Pp.377–384 in Paul A. M. van Lange (ed.) *Bridging Social Psychology: Benefits of Transdisciplinary Approaches* (2006) New York: Elbaum.

Williams, Terry, and Trevor B. Milton. *The Con Men: Hustling in New York City*. New York: Columbia University Press, 2015.

Wu, Junhui, Daniel Balliet, and Paul AM Van Lange. "Reputation, Gossip, and Human Cooperation." *Social and Personality Psychology Compass* 10 (6) (2016): 350–364.

Arnout van de Rijt and Vincenz Frey

7 Robustness of Reputation Cascades

Abstract: Reputation systems facilitate global exchange by allowing perfect strangers transacting across vast geographical distances to nonetheless trust one. However, a recent contribution by Frey and van de Rijt (2016, Scientific Reports) shows that reputation systems can produce "reputation cascades" whereby all trustors (e.g. all buyers) choose to transact with a single trustee (e.g. one seller), reinforcing the latter's monopoly on repute. This has the unintended consequence of generating winner-take-all inequality of an arbitrary nature, that is, between equally trustworthy trustees. The current study uses computer simulations to investigate the robustness of reputation cascades to noise. Results show that the dynamics of reputation formation continue to produce high levels of arbitrary inequality when information about the past behavior of trustees' is not always accurate and when trustors sometimes choose exchange partners randomly.

7.1 Introduction

Reputation systems are designed to mitigate *trust problems*. In a trust problem, mutually profitable exchange between two parties requires that one party – the *trustor* – first expose herself to the risk of exploitation by the other party – the *trustee* (Coleman 1990; Dasgupta 1988; Kreps 1990). For example, in online trade where a buyer does not know whether the seller will ship the product after receiving the payment. Game theory offers useful tools to model trust problems (Buskens, Frey, and Raub 2018) and we here use one of them – the standard Trust Game shown in Figure 7.1 – to make precise what we mean by a trust problem. If the trustor were to place trust, then the trustee would maximize profit by abusing trust, earning 100 instead of the 50 he would earn when honoring trust. Anticipating this, the trustor withholds trust, securing a payoff of 30 instead of the 0 she would earn if trust were abused. The consequence of individually rational behavior in the trust game is that trustor and trustee fail to exchange, earning only 30 each, where they could have earned 50 each, had they somehow achieved a 'smooth' exchange.

The trust problem may be overcome through use of a reputation system. In a reputation system, the trustor before making a choice can observe past actions of the trustee in interactions with other trustors, as reported by those trustors. The reputation system thus makes it possible for the trustor to *learn* about the trustee's inclination. The reputation system also gives the trustor some *control* over the trustee

Arnout van de Rijt, European University Institute and Utrecht University
Vincenz Frey, University of Groningen

https://doi.org/10.1515/9783110647495-007

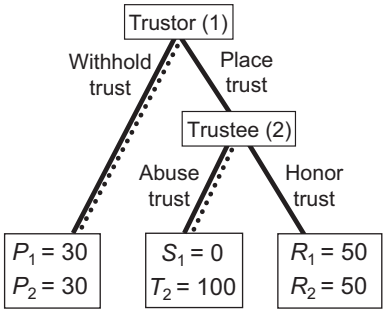


Figure 7.1: The standard trust game ($S_1 < P_1 < R_1$; $P_2 < R_2 < T_2$) with example values.

because the trustee has to take into account that the trustor could sanction bad behavior by leaving a negative rating, which would diminish the trustee's opportunity for future exchange. It may thus be rational for the trustee to honor trust, and hence for the trustor to place trust (Buskens 2002; Buskens and Raub 2002; Corten et al. 2016; Diekmann et al. 2014; Raub 1997).

Frey & van de Rijt (2016) identify an *unintended consequence* of reputation systems that has previously been overlooked. Unintended consequences result from the intentional actions of individuals which jointly aggregate – outside of any single actor's control – into collective behaviors no one wished for (Raub 2017; Raub, Buskens, and van Assen 2011; Raub and Voss 2016; Elster 1990; Boudon 1982). Specifically, reputation systems exhibit cumulative advantage (DiPrete and Eirich 2006; Merton 1968). If multiple trustees vie for the business of multiple trustors, then reputation systems will tend to produce extreme market concentration of an arbitrary nature. Namely, trustors will choose whichever trustee was lucky enough to be able to build up an initial reputation. As a result, this will set apart the chosen trustee's reputation even further, while other trustees are shunned. A *reputation cascade* (Frey and van de Rijt 2016) unfolds, propelling a single market leader. The market leader may not be any more trustworthy than many others who are never trusted and may even offer a worse deal than untried competitors. The distinction is largely random. This arbitrary inequality is the unintended consequence of a process of cumulative advantage.

Here we probe the robustness of reputation cascades. Frey & van de Rijt (2016) propose a game-theoretic model and demonstrate that in this model reputation cascades are an equilibrium. The behavior of participants in their laboratory experiment closely approximates this equilibrium, and – at the macro level – it does indeed exhibit reputation cascades. However, the external validity is not beyond doubt: the Frey & van de Rijt (2016) theoretic model makes some restrictive assumptions and their laboratory setting limits factors that could undermine cascading in "real-world" reputation systems. In particular, two assumptions in their theory and experiment are arguably often violated in real-world markets.

First, Frey and van de Rijt's (2016) experiment implemented automated, truthful feedback. Yet real-world systems are noisy. Trustees might fail to deliver on time for reasons out of their control. A timely shipped item might arrive late or even get lost. Such errors will result in negative ratings, disgracing generally reliable trustees. Some ratings may not reflect performance at all: Raters may err by giving a bad score when they mean to give a good one, or trustees may deceptively award themselves positive or their competitors negative ratings (for example, Diekmann et al. 2014; Luca and Zervas 2016). Real world systems may also exhibit noise in the trustors' selection of trustees: trustors may not always trust the trustee with the best reputation. The absence of search costs or other aspects of differentiation among trustees probably minimized such noise in trustor behavior in the Frey and van de Rijt (2016) experiment. Do reputation cascades continue to happen under such noisy conditions?

Second, the original experiment cannot tell us about the sustainability of reputation cascades, as it had relatively short interaction sequences. Empirical findings suggest decreasing marginal effects of ratings on trust (Przepiorka, Norbutas, and Corten 2017). If the marginal impact of positive ratings were decreasing, limiting the advantage of highly reputable trustees over those lacking a reputation, would reputation cascades break down as initially disadvantaged trustees catch up? In the present article, we address these questions using computer simulations.

7.2 The model – A market with trust problems and a reputation system

In this section, we describe the model that we subsequently analyze game-theoretically (Section 7.3) and in computer simulations (Section 7.4). It is a modified version of the model introduced by Frey & van de Rijt (2016).

We assume an indefinitely large population of trustors and $N > 1$ trustees. Each round 1, 2, 3, . . ., one trustor is chosen to play, and no trustor plays more than once. After every round, the next round is played with probability $0 < w < 1$ while the game ends with probability $1 - w$.

Every round, the trustor i chosen to play decides whether to withhold trust or to select one of the trustees and place trust in that trustee. If i does not place trust in any trustee, i receives payoff P_1. If i chooses to place trust in trustee j, then j can honor or abuse trust. Honored trust earns i and j the payoffs $R_1 > P_1$ and $R_2 > P_2$, respectively. If trust is abused, i's and j's payoffs are $S_1 < P_1$ and $T_{2j} > R_2$, respectively. Any trustee who is not selected by i receives payoff P_2. We do not need to make an assumption on what the payoffs in a round are for trustors who are not at play, and we assume no discounting of payoffs from future rounds (apart from the weighting related to the continuation probability w).

Trustees differ in the payoff T_{2j} that they can earn when abusing trust. T_{2j} is drawn independently for each trustee j before round 1 from a probability distribution with unbounded density F. While F is common knowledge, the actual manifestation of T_{2j} is private information of trustee j. As we will see later (Lemma 2), the assumption of heterogeneity in T_2 implies that one trustee may not be trustworthy under precisely the same circumstances that incentivize another trustee to honor trust, which is empirically plausible and crucial in the formation of reputation cascades.

A reputation system makes the entire history of all trustees' choices available to all trustors. Our game-theoretic analysis assumes that this reputation information is always truthful, and we will relax this assumption in the simulation analysis.

7.3 Game-theoretic analysis

In the game-theoretic analysis, we first postulate a strategy for trustors (Lemma 1). Loosely speaking, this strategy is to place trust in any unknown trustees as long as there is no trustee with a good reputation and otherwise to place trust in any trustee with a good reputation. It can be shown that this strategy is a best-response strategy for trustors – that is, a trustor cannot improve her expected payoff by playing a different strategy – if (*i*) each trustee either always honors trust or always abuses trust and (*ii*) the portion of trustworthy trustees is "large enough". Lemma 2 shows that if trustors play that strategy, then indeed the best-response strategy for some trustees is to always honor trust, while all other trustees always abuse trust (the prospect of being trusted again makes trustees with a low T_{2j} trustworthy but it is not a sufficient incentive for trustees with a high T_{2j} to honor trust). Finally, Proposition 1 states the condition for an equilibrium in which the trustors play the strategy defined in Lemma 1 and some portion trustees honor trust, which includes the specification of the critical "large enough" proportion of trustworthy trustees. We conclude by discussing the emergence of arbitrary inequality in this equilibrium. This small game-theoretic analysis reproduces the analysis by Frey & van de Rijt (2016) for a simplified scenario, and we refer the reader to that paper for the proofs of our Proposition and Lemmas.[1]

If each trustee is either trustworthy or untrustworthy throughout, we can distinguish three sorts of trustees based on their history: (1) trustees who honored trust in some past rounds, (2) trustees who abused trust in some past rounds, and (3) trustees who were never trusted. Using this distinction and given the assumption that

1 For the proof of Lemma 1 and Proposition 1, see the proof of Proposition 1 in Frey & van de Rijt (2016). For the proof of Lemma 2 of the current paper, see the proof of Lemma 1 in Frey & van de Rijt (2016).

the portion of trustworthy trustees is large enough (in a sense to be specified in Proposition 1), the best-response strategy of a trustor *i* selected to play is:

Lemma 1 – Strategy of a trustor *i*.

- *If there are trustees who honored trust in some past rounds, i places trust in one of them (randomly choosing one).*
- *If there are no trustees who honored trust in some past rounds but there are trustees who were never trusted, i places trust in one of these trustees (randomly choosing one).*
- *If there are no trustees who honored trust and also no trustees who were never trusted, i withholds trust.*

If trustors behave as specified in Lemma 1, game-theoretic rationality implies that a trustee with a sufficiently low T_{2j} always honors trust while trustees with a high T_{2j} always abuse trust:

Lemma 2 – Strategy of a trustee *j*. *A trustee j's best response to the strategy of trustors defined in Lemma 1 is to honor trust when selected if and only if*

$$w \geq \frac{T_{2j} - R_2}{T_{2j} - P_2} \Leftrightarrow T_{2j} \leq \frac{R_2 - wP_2}{1 - w} \tag{7.1}$$

So a proportion ρ of trustees always honor trust when selected, where

$$\rho = \int_0^{\frac{R_2 - wP_2}{1-w}} dT_{2j} \tag{7.2}$$

It is only rational for a trustor to play the strategy in Lemma 1 if the portion ρ of trustworthy trustees is so large that compared to no trust, trustors prefer taking the risk of placing trust in a trustee who does not yet have any reputation.

Proposition 1. *There exists a Nash equilibrium in which all trustors play the strategy defined in Lemma 1 and some portion ρ of trustees honor trust when selected if and only if*

$$\rho \geq \frac{P_1 - S_1}{R_1 - S_1} \tag{7.3}$$

Consider now how this equilibrium implies the emergence of arbitrary inequality among trustworthy trustees (trustees for which $T_{2j} \leq (R_2 - wP_2)/\ (1-w)$). The strategy of trustors implies that the first of these trustees who gets the chance to honor trust will be trusted in all future rounds. Intuitively, staying with the first established trustee allows the trustors to avoid the risk of trust abuse. Thus, this small game-theoretic analysis implies the emergence of extreme inequality among equally

trustworthy trustees – one trustworthy trustee gets all the business while all other trustworthy trustees walk empty-handed. In the next section we replicate this result in a computational model and investigate its sensitivity to noise in trustor and trustee behavior and to decreasing marginal returns of reputational information.

7.4 Simulation analysis

7.4.1 Simulation approach

In the simulations we introduce noise and decreasing marginal returns into the model described in Section 7.2 as follows:
Trustee noise: We assume that – as in the game-theoretic equilibrium – some trustees are trustworthy, always honoring trust, while all other trustees are untrustworthy, always abusing trust. However, these actions are subject to noise. With probability α a trustworthy trustee's action is reported as abuse (either because the trustee made a mistake or because the trustee's action was reported wrongly). Likewise, with probability α an untrustworthy trustee's action is reported as honoring trust.

Trustor noise: We continue to assume that trustors behave as stated in Lemma 1 – exhibiting a tendency to choose trustees with a good reputation, but do so probabilistically, following a logistic choice function:

$$p(J=0) = \frac{1}{1+\sum_k e^{\beta\left(g_k{}^s - 4b_k{}^s\right)}} \tag{7.4}$$

$$p(J=j) = \frac{e^{\beta\left(g_j{}^s - 4b_j{}^s\right)}}{1+\sum_k e^{\beta\left(g_k{}^s - 4b_k{}^s\right)}} \tag{7.5}$$

Here, $J = 0$ represents the choice to withhold trust, $J = j$ the choice to trust trustee j, g_j and b_j the numbers of *g*ood respectively *b*ad ratings given to trustee j, coefficient β representing the degree to which trustee choice is guided by reputational information instead of randomness. A larger β implies less randomness in the trustors' choices of trustees.

Marginal returns of ratings: This is controlled by parameter $s \in [0,1]$, with $s = 1$ implying a constant impact of every additional rating on the chance of a trustee being selected.

Throughout, model parameters that are not of interest are fixed at what we judge to be reasonable values. A full exploration of the entire parameter space including these parameters is beyond the scope of this chapter. Table 7.1 shows the parameters that are fixed with their respective values. By fixing the number of rounds to be equal to the number of trustors (N_1 = 100,000) we ensure that each

Table 7.1: Parameters fixed in simulations.

Parameter	Fixed at
Number of rounds in a game	100,000
Number of trustors (N_1)	100,000
Number of trustees (N_2)	100
% trustworthy trustees	50%
Importance of negative vs. positive ratings	4x

trustor moves only once. By keeping the number of trustees ($N_2 = 100$) and the ratio of trustors to trustees (1000:1) high, we ensure that a random distribution of trust produces minimal inequality.

The parameters of interest that we vary are noise in trustee reputational information (due to either trustors reporting wrongly, or trustee mistakes), α, sensitivity to reputational info vs. noise, β, and the marginal impact of reputational information, s.

We quantify the degree of arbitrary inequality as the Gini coefficient calculated among trustworthy trustees. Any distinction made between trustworthy and untrustworthy trustees is justified and non-arbitrary, which is why untrustworthy trustees' market shares must be excluded from inequality assessment. The Gini coefficient equals half the average difference in market share and ranges from $2/N_2$ (equal market shares) to $1–2/N_2$ (monopoly).

7.4.2 Simulation results

Figure 7.2 shows the level of arbitrary inequality over rounds averaged across 100 simulation runs, for different noise levels. The marginal impact of ratings is assumed to have a fixed level of $s = 1$. Trustee noise α and trustor noise β are simultaneously varied at a joint rate of $r = 2\alpha = 2 / (1 + e^{\beta})$. When noise is absent ($r = 0$), trustor behavior is similar to the strategy defined in Lemma 1: Initially, trustors place trust in trustees who were never trusted and after the first trustworthy trustee is found, all subsequent trustors place trust in this trustee (who keeps honoring trust). This is reflected in the arbitrary inequality level shooting up to its maximum and staying there, showing the predicted enduring monopoly situation. Thus, if noise is absent, the simulation reproduces the result of the game-theoretic analysis.

Figure 7.2 further shows that when noise is introduced – trustors occasionally deviate randomly and trustworthy trustees occasionally receive negative ratings – arbitrary inequality continues to be high. At each noise level, arbitrary inequality approaches a stable level. When noise is increased to 40%, this stable level of

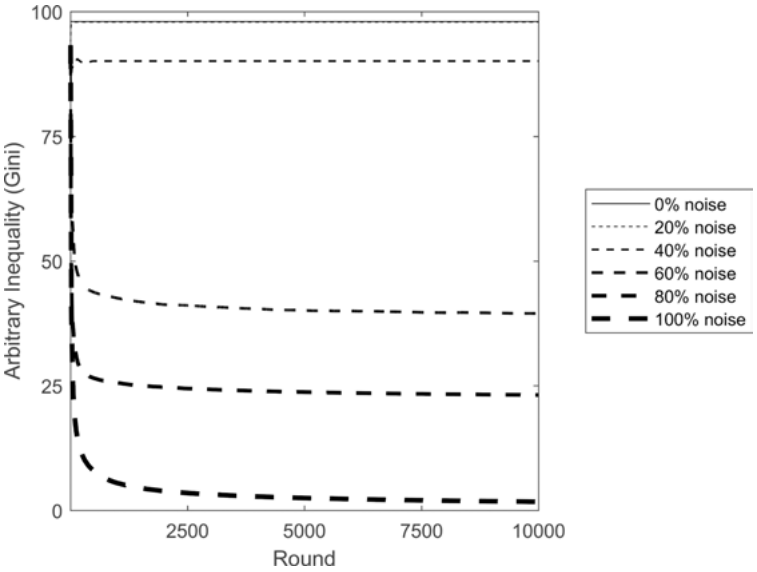


Figure 7.2: Levels of arbitrary inequality under various levels of noise, averaged across 100 simulation runs.

inequality is still near-maximal (Gini = 92%), demonstrating strong robustness of the original result. When noise is increased further, inequality drops more rapidly. At 60% noise ($\alpha = (1 + e^{\beta})^{-1} = .3$; i.e. $r = .6$) – when trustors act randomly more than half the time when they choose between two trustees who differ by 1 positive rating, and when trustees' actions receive random ratings 60% of the time – arbitrary inequality has become moderate (Gini = 41%). Higher noise levels produce little to no inequality.

Table 7.2 explores robustness of arbitrary inequality formation when trustor noise and trustee noise are independently varied (while in Figure 7.2, trustor noise and trustee noise were simultaneously varied). The values in Table 7.2 correspond to the inequality levels in the last round, at the right vertical axis of Figure 7.2. Table 7.2 shows that the emergence of arbitrary inequality is more sensitive to trustee noise: When moving up in Table 7.2, inequality levels drop more quickly than when moving left. Even at very high levels of trustor noise, inequality is near-maximal, as long as trustee noise is modest (bottom left of Table 7.2). This happens because with high trustor noise, even though a trustor's choice between trustees with mildly differing reputations is now practically a coin flip, a very obvious choice such as between a market leader and a competitor who has never been trusted before is still often made correctly. This allows emergent market leaders to extend their lead even as other trustees are regularly given a chance. This suggests that the assumption of constant returns – which we will relax later – is crucial for the ability of market leaders to maintain their dominance at high noise levels.

When instead trustor noise levels are kept low while trustee noise is increased (top right of Table 7.2), levels of inequality do drop. At 100% trustee noise, when reputation systems are entirely dysfunctional, inequality levels remain around 40%. This happens because the first cohorts of trustors act on the useless information they generate for one another, choosing popular trustees until they are reported to have abused trust (50% of the time). Later generations gradually retreat to the choice not to trust anyone, leaving inequality in this failed market at moderate levels.

Table 7.2: Arbitrary inequality (Gini) in final round, by trustor noise (hor.) and trustee noise (vert.), under *constant* marginal impact of ratings ($s = 1$), averaged across 100 simulation runs.

		Noise in trustors' selection of trustees ($2/[1+e^{\beta}]$)										
		100%	**90%**	**80%**	**70%**	**60%**	**50%**	**40%**	**30%**	**20%**	**10%**	**0%**
Noise in trustee reputation (2α)	100%	2	14	19	23	26	29	30	34	34	39	39
	90%	2	16	21	25	28	31	33	36	37	42	45
	80%	2	17	23	27	31	35	36	40	41	45	47
	70%	2	20	26	30	34	37	40	43	45	49	54
	60%	2	23	30	35	39	42	45	49	51	55	61
	50%	2	30	39	44	49	52	55	59	61	63	78
	40%	2	85	88	90	90	90	91	93	91	92	93
	30%	2	95	96	97	97	97	97	97	98	97	97
	20%	2	96	97	97	98	98	98	98	98	98	98
	10%	2	97	98	98	98	98	98	98	98	98	98
	0%	2	98	98	98	98	98	98	98	98	98	98

Finally, in analysis summarized in Table 7.3, we relax the common assumption of constant marginal returns of success ($s = 1$) in models of positive feedback (Allison 1980; Barabási and Albert 1999; Coleman 1964; DiPrete and Eirich 2006; Price 1976; Simon 1955), exploring reputation dynamics when additional ratings are worth less and less ($s = ¾, ½, ¼, 0$). Strikingly, in the absence of noise, inequality remains at its maximal level (Gini = 98%) even as s is decreased to very low levels (¼), where the marginal value of extra ratings is sharply diminishing. As one may expect, as we increase trustor and trustee noise simultaneously, inequality levels drop more rapidly when s is also decreased. Still, even under these more stringent conditions, significant market concentration obtains at modest noise levels.

Table 7.3: Arbitrary inequality in final round, by noise (r) and marginal impact of ratings (s = 0, ¼, ½, ¾, 1), averaged across 100 simulation runs.

s r	100%	90%	80%	70%	60%	50%	40%	30%	20%	10%	0%
0	2	2	2	2	2	2	2	2	2	2	2
¼	2	2	3	4	6	9	13	17	25	37	98
½	2	5	9	12	17	22	27	34	44	98	98
¾	2	10	16	22	28	35	44	67	98	98	98
1	2	16	23	30	39	52	92	97	98	98	98

7.5 Discussion

In sum, reputation cascades exhibit reasonable robustness along the dimensions explored here. In markets with trust problems, reputation systems create market concentration. This happens because of positive feedback perpetuating the advantage of established parties of good repute. We have shown that this theoretical mechanism is not overly sensitive to behavioral deviations. A small number of arbitrarily selected trustees continue to dominate in market share when these established trustees occasionally get bad ratings, and also when buyers occasionally select unknown trustees lacking a reputation. Reputation cascades also persist in settings with mildly decreasing marginal benefits of additional ratings, but do cease to occur when marginal benefits are strongly decreasing.

The paper does not exhaust all relevant dimensions along which robustness could be assessed. One interesting extension of the model is one where outdated ratings do not count as much. This is applicable in tumultuous scenarios where businesses are frequently passed on from one owner to another, and the new management may differ in trustworthiness, or in settings where external factors change trustworthiness motives. Under such regimes, it may be easier for someone lacking a reputation to establish one, as long-established parties cannot indefinitely enjoy their yesteryear advantages.

Another natural extension is the addition of a price mechanism: What if excluded sellers could combat reputable sellers by undercutting their bids? We conjecture that under these circumstances reputation cascades will continue to obtain. An equilibrium should be possible in which the winning trustee asks a price that exceeds competitors' prices by an amount – a "reputation premium" – (Diekmann et al. 2014; Przepiorka, Norbutas, and Corten 2017; Snijders and Weesie 2009; Przepiorka & Aksoy 2020) that is small enough to withhold most buyers from exploring other options, keeping them loyal. Implementing such a price mechanism would, however, add considerable

complexity to our simulation approach, requiring one to specify the trustees' pricing strategy as well as how trustors trade off prices against reputations. Consistent with these predictions, empirical studies on reputation premiums document that many sellers do indeed *not* manage to attract any customers. In Snijders & Weesie's (2009: 175) study on an online programming market, 82% of the programmers who offered their services were never selected by even a single customer while the remaining 18% sold a total of 22,506 programming jobs in the observation period (see also Barwick and Pathak 2015; Diekmann et al. 2014).

Finally, one may ask whether the tendency of reputation systems to induce herding on a few established trustees has further negative implications, in addition to creating baseless market concentration. Can lock-in on a trustee with an established reputation prevent trustors from exchanging with an excluded trustee with a better value proposition? In the present model, only trustworthy and untrustworthy trustees exist, so that market concentration always exists around trustworthy trustees. A possible extension of the model allows three or more types who differ in the value they provide to trustors, allowing an exploration of the possibility of inferior lock-in on a trustee who occasionally abuses trust or delivers less value when honoring trust. Such lock-in may be likely when every trustor buys just once – as we assumed in this chapter – and it may be less likely when the same trustors buy several times – as assumed in the original model of Frey & van de Rijt (2016) – because trustors can then benefit from their own exploration efforts.

All in all, the present investigation confirms the emergence of arbitrary inequality in markets as a robust phenomenon. As trustors flock to safe havens, untried exchange alternatives that may be qualitatively equivalent, or even superior, are left unexploited.

References

Allison, Paul D. 1980. "Estimation and Testing for a Markov Model of Reinforcement." *Sociological Methods & Research* 8 (4): 434–453.

Barabási, Albert-László, and Réka Albert. 1999. "Emergence of Scaling in Random Networks." *Science* 286 (5439): 509–512.

Barwick, Panle Jia, and Parag A. Pathak. 2015. "The Costs of Free Entry: An Empirical Study of Real Estate Agents in Greater Boston." *The RAND Journal of Economics* 46 (1): 103–145.

Boudon, Raymond. 1982. *The Unintended Consequences of Social Action. New York: St.* Martin's Press.

Buskens, Vincent. 2002. *Social Networks and Trust*. Boston, MA: Kluwer.

Buskens, Vincent, Vincenz Frey, and Werner Raub. 2018. "Trust Games: Game-theoretic Approaches to Embedded Trust." In *The Oxford Handbook of Social and Political Trust*, ed. Eric M. Uslaner, 305–331. Oxford: Oxford University Press.

Buskens, Vincent, and Werner Raub. 2002. "Embedded Trust: Control and Learning." *Advances in Group Processes* 19, 167–202.

Coleman, James S. 1964. *Introduction to Mathematical Sociology*. New York: Free Press of Glencoe.

Coleman, James S. 1990. *Foundations of Social Theory*. Cambridge, MA: Belknap Press of Harvard University Press.

Corten, Rense, Stephanie Rosenkranz, Vincent Buskens, and Karen S. Cook. 2016. "Reputation Effects in Social Networks Do Not Promote Cooperation: An Experimental Test of the Raub & Weesie Model." *PloS One* 11 (7): e0155703.

Dasgupta, Partha. 1988. "Trust as a Commodity." Pp. 49–72 in *Trust: Making and Breaking Cooperative Relations*, ed. Diego Gambetta. Oxford: Blackwell.

Diekmann, Andreas, Ben Jann, Wojtek Przepiorka, and Stefan Wehrli. 2014. "Reputation Formation and the Evolution of Cooperation in Anonymous Online Markets." *American Sociological Review* 79 (1): 65–85.

DiPrete, Thomas A., and Gregory M. Eirich. 2006. "Cumulative Advantage as a Mechanism for Inequality: A Review of Theoretical and Empirical Developments." *Annual Review of Sociology* 32: 271–297.

Elster, Jon. 1990. "Merton's Functionalism and the Unintended Consequences of Action." Pp. 129–135 in *Robert K. Merton: Consensus and Controversy*, ed. Jon Clark, Celia Modgil and Sohan Modgil, 129–135. London: The Falmer Group (Taylor and Francis).

Frey, Vincenz, and Arnout van de Rijt. 2016. "Arbitrary Inequality in Reputation Systems." *Scientific Reports* 6: 38304.

Kreps, David M. 1990. *Game Theory and Economic Modelling*. Oxford: Oxford University Press.

Luca, Michael, and Georgios Zervas. 2016. "Fake It till You Make It: Reputation, Competition, and Yelp Review Fraud." *Management Science* 62 (12): 3412–3427.

Merton, Robert K. 1968. "The Matthew Effect in Science: The Reward and Communication Systems of Science Are Considered." *Science* 159 (3810): 56–63.

Price, Derek de Solla. 1976. "A General Theory of Bibliometric and Other Cumulative Advantage Processes." *Journal of the Association for Information Science and Technology* 27 (5): 292–306.

Przepiorka, Wojtek, and Ozan Aksoy. 2020. "Does Herding Undermine the Trust Enhancing Effect of Reputation? An Empirical Investigation with Online-Auction Data." *Social Forces*, in press.

Przepiorka, Wojtek, Lukas Norbutas, and Rense Corten. 2017. "Order without Law: Reputation Promotes Cooperation in a Cryptomarket for Illegal Drugs." *European Sociological Review* 33 (6): 752–764.

Raub, Werner. 1997. *Samenwerking in Duurzame Relaties En Sociale Cohesie (Cooperation in Durable Relations and Social Cohesion)*. Amsterdam: Thesis Publishers.

Raub, Werner. 2017. *Rational Models*. Utrecht: Utrecht University.

Raub, Werner, Vincent Buskens, and Marcel A. L. M. van Assen. 2011. "Micro-Macro Links and Microfoundations in Sociology." *The Journal of Mathematical Sociology* 35 (1–3): 1–25.

Raub, Werner, and Thomas Voss. 2016. *Micro-Macro Models in Sociology: Antecedents of Coleman's Diagram*. Pp. 11–36 in *Social* Dilemmas, *Institutions and the Evolution of Cooperation*, eds. Wojtek Przepiorka and Ben Jann. Berlin: De Gruyter.

Simon, Herbert A. 1955. "On a Class of Skew Distribution Functions." *Biometrika* 42 (3–4): 425–40.

Snijders, Chris, and Jeroen Weesie. 2009. "Online Programming Markets." Pp. 166–185 in *eTrust: Forming Relationships in the Online World*, eds. Karen S. Cook, Chris Snijders, Vincent Buskens and Coye Cheshire. New York: Russell Sage.

Henk Flap and Wout Ultee

8 Organized Distrust: If it is there and that Effective, Why Three Recent Scandals?

Abstract: In the field of society and politics, Popper replaced the question of who should rule by that of how to reduce damages rulers impose upon their subjects. Popper's answer was organized distrust, institutions like periodic elections for public posts and the separation of powers. By detailing three scandals that played out in the 2010s, this contribution asks to what extent organized distrust was effective in these instances. The first case is that of fraud by the Dutch social psychologist Stapel. Our conclusion is that serial offenders are not likely to be caught by the much-praised double blind review system. The second scandal is deceptive software in cars for monitoring emissions. It came about through three causes. First, the European Union set emission standards, but left surveillance to national governments, making for Hobbes's covenants without swords. Second, national organizations for applied research did find that emissions were too high, but were not allowed to publicize this, making for Durkheimian fatalism. Third, executives of car companies set impossible goals for engineers, making for Mertonian anomie in the form of fiddling. The third scandal is sexual abuse of minors by Catholic clergy. Catholic authorities settled cases behind closed doors, lowering the number of trials because victims did not have models for complaints. All in all, Popper's methodological individualism should allow for corporate actors next to individual actors, so that Coleman's proposition of increasing asymmetries in contemporary societies may be tested.

8.1 Personal trust and organized distrust

Trust cements societies, confidence of people in each other lubricates life. For the systematic study of societies, these avowals are a bit metaphorical. Yet an issue lurks here, and it becomes visible by spelling out corollaries of the thesis that *peaceful coexistence* is impossible without trust. To begin with: in societies rife with *dis*trust, life will be solitary, poor, nasty, brutish and short – that is the old quote from *Leviathan* (Hobbes 1651). In addition: societies where doubts abound will *Collapse* – to borrow a recent book title (Diamond 2005). All this amounts to a major thesis in current sociology.

Yet there is a competing conjecture. In societies with the rule of law, free labor, elected politicians, and similar arrangements, unanticipated developments will

Henk Flap, Utrecht University
Wout Ultee, Radboud University

https://doi.org/10.1515/9783110647495-008

occur like Smith's accumulation of stocks and rising real recompenses for labor, and Millar's less unequal balances of the rights of husbands and wives, fathers and children, masters and servants, as well as rulers and subjects. These provisions also promote processes like Durkheim's bonding between a society's members, and Weber's rationalization of economies, polities, and other areas of life. Societies with these arrangements institutionalize criticism and organize distrust, occasioning long-term trends studied by sociology's Scottish ancestors and its Continental initiators. So, there is the issue of trust and distrust. In this contribution we seek to deal with it, and we do so by way of a data driven essay.

8.1.1 A problem shift in studies of societies and politics

Popper argued – in *The Open Society and Its Enemies* from 1945, and more extensively in its 1952 edition with an augmented chapter entitled "The principle of leadership" – that theorizing about societies and their polities has been dominated since two millennia by Plato's ill-considered question of who shall come to rule a society's state. Rather, the pivotal question should be how to eliminate bad rulers without bloodshed, and how to reduce damages rulers impose on their subjects (Popper 1952: i, 121). This problem shift puts proposals for political reform in a different light, unifies them, and suggests additional amendments. Even if heads of states are benevolent and wise upon ascension, they – following Acton – may be corrupted by their exclusive rights to draft men as soldiers, tax households, convict persons who trespass rules protecting life and property, and issue travel documents. By voting in elections, people express trust in persons running for office. If leaders are elected for life, the impact of good ones who turn bad may be large. However, if laws demand that heads of states periodically face re-election, people may voice distrust by voting for their opponents in the next election. And if – as Bentham proposed – every subject is entitled to vote, harm will be even smaller. Other arrangements reduce the injurious impact of rulers by limiting their powers, and by balancing the curbed powers against other powers. Montesquieu's *trias politica* does away with the concentration of powers in one office held by one person, and amounts to another case of 'organized distrust'. Popper did not use this expression, but it nicely sums up his argument.

Popper criticized his own theory that organized distrust lowers violence between rulers and subjects by maintaining that institutions are like fortresses. They should be well-designed – organized distrust would improve an institution's design. They should be well-staffed too – but Popper did not enter into that matter. So, how to fill that hole? We try to do so in this contribution.

Given the problem shift proposed by Popper for research on societies and their polities, the issue is not about the amount of trust of persons in one another, and the many data around on the people's trust in their government, the police, and the

courts, are irrelevant for Popper's new problem. This remains so if trust of persons in – say – banks, multinationals, and universities is ascertained. Following Popper, the consequential question is about non-personal entities distrusted or trusted by people, and the extent to which they contain arrangements allowing for and inviting comments, examination, and inspection. Some we mentioned, others include reports on funds raised and spent, and their certification by auditors. A complete list of provisions for criticism seems unattainable, as new stipulations may crop up after scandals, and a world without scandals will stay far away. That is why we conclude our dissection of three recent scandals with proposals for more organized distrust.

Accepting Popper's problem shift, it is wise in sociology to distinguish between on the one hand natural actors like you as a reader and the authors of this essay, and on the other hand states, or more generally corporate actors. This is a common distinction in jurisprudence. Of course, natural persons hold the political offices of states, but these natural persons have, given their formal position, additional capacities to act. The present contribution will not deal with the amount of trust of natural persons in corporate actors, but the way in which distrust of corporate actors is organized in contemporary societies like the Netherlands. This paper bypasses questions about organized distrust within their polities, and seeks to answer questions about organized distrust for three other corporate actors, to wit editorial boards of scientific periodicals, car companies, as well as churches. If Popper's theory of organized distrust is to demonstrate its mettle, it does so by its application to other types of corporate actors than states and their polities, for which it was developed.

8.1.2 Institutional individualism as a theoretical method for sociology?

In *The Open Society* Popper also argued in favor of theories with a specific content for sociology in general. These arguments have created quite some confusion. Popper asserted that sociology is autonomous, but rejected methodological collectivism, backed methodological individualism, at the same time opposing psychologism. It does not seem easy to square these proposals. We sort things out, and the outcome will be that the common distinction between hypotheses about collectivities and hypotheses about individuals – as featured in the debate in the United Kingdom assembled by O'Neill (1973) in *Modes of Individualism and Collectivism* – should be replaced by that between hypotheses about corporate actors and hypotheses about natural persons, and supplemented by the recognition that sociology's questions involve a combination of natural and corporate actors, as well as by explanations of societal phenomena that invoke corporate actors next to individual actors.

In "The autonomy of sociology", a chapter of *The Open Society*, Popper pleaded for a sociology bypassing explanations that invoke human nature and its instincts. That would amount to sterile *psychologism*. According to Popper, it would be wrong to argue that the conventions of social life are outcomes of motives springing from the mind of individual persons. If motives are to explain societal phenomena, the explanation also should refer to the environment in which people live, in particular the institutions of their social setting. Popper then does not announce what should be trumpeted (ii, 90): the outcomes of human actions not only depend on the *motives*, but also on the *opportunities* of persons. Anyway, the term methodological *individualism* is unfortunate, because explanations adducing solely the motives of persons – or more generally human nature – are individualist too. This is perhaps why some scholars came to favor the tag institutional individualism, such as the philosopher Agassi in *The British Journal of Sociology* from 1974. We replace the term psychologism by the expressions instinctual individualism and motivational individualism.

What exactly is wrong, according to Popper, with instinctual and motivational individualism? One of Popper's examples involves wealth. John Stuart Mill had maintained in 1843 that a finding to the effect of most goods being produced by markets, in the end can be derived from the tendency residing in human beings to pursue wealth. Popper disagrees by arguing almost – but not exactly and less overtly – like Smith, the founder of economics. Nations differ in their standard of living, how is this possible if all people had, have and will have the same propensity to improve their life? To answer this question, Smith (1776) in *The Wealth of Nations* did not coolly claim that this inclination differs in intensity from country to country. In some areas of the world people live more closely together than in other parts, and it is easier to transport goods on navigable rivers and seas than over hilly and rugged lands. Markets differ in size, and when markets are larger, work in factories become more divided, and the division of labor, in its turn, makes for cheaper products, higher profits, and an increase in a nation's wealth. Some settings offer definitely more chances for improving one's living standard than other circumstances, and the freedom of enterprise heralded by Smith augments them. It now is clear what the difference between Popper's psychologism and individualism looks like. Some explanations invoke motives only, other accounts adduce motives plus opportunities. Later we discuss the *methodological* of Popper's individualism.

Incest is Popper's counterexample to instinctual individualism. To explain why every society forbids incest, it would not help to note that every young man everywhere goes out of his head when asked how often he has sex with his sister. Rather, the stronger a society's sanctions against incest, the deeper its elders plant in juveniles the idea that incest is dead wrong, and as a consequence almost everyone everywhere is disgusted by it. Here it is a pity that Popper does not scrutinize the debate around 1900 between Westermarck and Durkheim. Westermarck held the first sociology chair in Europe, and postulated an instinct against incest, evolved by

natural selection. Durkheim took avoidance of certain marriages as the other side of rules promoting exogamy, and the taboo on incest in societies like his native France as a survival of the injunction – still present among Australian and North American aboriginals – to marry someone with a different totem. Durkheim hinted that this rule would decrease hostility between totem groups, and fulfill a societal need for peace.

But then – since persons may die of hunger and thirst – how is the societal need for peace met? In "La Prohibition de l'Inceste et ses Origines", Durkheim (1898) held that among hunters and gatherers the men with one totem exchange women of their own totem against women of another totem offered by men of that totem, with peace as a byproduct. Lévi-Strauss (1949) developed this hint in *Les Structures Élémentaires de la Parenté* into a full-fledged theory. Durkheim also showed that among hunters and gatherers in Australia, adolescent men not only eschew sex with women of their own totem. Men and women of the same totem do not share meals, and as unmarried adults they avoid one another during leisure. The question of why no incest was a sub-problem of that of avoidance between persons of the same totem and different sex. O'Neill (1973) neglected Durkheim on incest.

Now a case of purported instinctual individualism that boils down to a case of budding institutional individualism, bringing out the issue of whether it is logically possible to explain societal differences by human nature. Hume, the infidel and friend of professor Smith (we allude to Rasmussen's (2017) disappointing book about a friendship between two Scots who shaped modern thought), raised in *The Natural History of Religion* the question of the origin of religion, immediately turning it into that of its origin in human nature (Hume 1757). All the same, Hume had misgivings about the question of religion's roots in human nature. He accepts that beliefs in several co-existing invisible and intelligent powers prevailed all over the world until (then) 17 centuries ago. However, since that time the idea of one such power spread all over Europe. In addition, no two nations with polytheism ever fully agreed in religious matters. What is more, some nations do not have a religion. All this does not square easily – so Hume notes – with the idea of human instincts the same in all times at all places.

How does Hume pull off instinctual individualism? He does not, and spins instinctual into institutional individualism. Hume winds up his *History* by stating that '(t)he universal propensity to believe in invisible, intelligent power, if not an original instinct, . . . at least (is) a general attendant of human nature'. Even if one were to know what amounts to a general attendant of human nature, it does fall outside the scope of instinctual individualism, and when the content of Hume's specific hypotheses is considered, Hume strays even more from it. In his discussion of polytheism Hume posits that if for persons visible factors do not do the job of explanation, they will postulate invisible and intelligent powers: the vicissitudes of human lives – in particular sickness, want, and adversity – suggest many powers influencing

human weal and woe. Here – in our reading – Hume explains polytheism by appealing to the natural plus social environment. Also, the members of a society 'reduce heavenly objects to the model of things below', and out of several gods they 'may represent one god as the prince or supreme magistrate of the rest, who . . . rules them with an authority, like that which an earthly sovereign exercises over his subjects and vassals'. In this account of monotheism – so we hold – Hume invokes an institution, and notably a corporate actor *absent* among hunters and gatherers. It is an earthly sovereign, and occasions in the mind of a sovereign's subjects the idea of a heavenly sovereign. Finally, Hume takes his hypothesis about life's vicissitudes as corroborated by noting that in his days gamblers and sailors are the most superstitious persons in Scotland. So, the people most vulnerable to capricious fortune, adduce the longest list of whimsical entities to account for how they fare. Two and a half centuries ago the division of labor did not comprise sports super stars.

Of course, Hume was on the right track when explaining a society's religion with the hypothesis that its members take invisible factors as resembling visible entities like human beings endowed with intelligence and will. All the same, Durkheim (1912) paved a way forward with the thesis that religions draw on societal analogies. He held in *Les Formes Élémentaires de la Vie Religieuse* that among Australian hunters and gatherers such sociomorphism was common, and that only in later stages of what now is called technological development, the tendency evolved to account for phenomena by willful and wise invisible entities. This anthropomorphism therefore cannot be an instinct and part of human nature, it is acquired under societal conditions. Topitsch (1958) sorted things out in *Vom Ursprung und Ende der Metaphysik*: people comprehend the faraway and unknown by analogy with the known and nearby, implying that to the extent people live in different settings, their thoughts about the distant and unfamiliar vary. With the cognitive and evolutionary turns in psychology, this principle gained ground. In Pinker (2002) thinking in metaphors is a human universal. However, for Pinker – ignoring Durkheim – anthropomorphization is a human universal too.

Now, what exactly is wrong, according Popper, with collectivism? Popper contrasted individualism with collectivism, and took Durkheim (1895) – who pleaded in *Les Règles de la Méthode Sociologique* for an autonomous sociology – as a proponent of collectivism. Durkheim definitely favors seeking causes of societal phenomena among other societal phenomena, and he found them there too. According to Popper, in a sentence added in 1952 to the chapter entitled "The open society and its enemies" in his book with the same title, Durkheim's methodological rule was that societies must be analyzed in terms of concrete social groups (i, 175). Popper's examples of concrete groups are persons related by blood, people who live at the same place, human beings who shared a common effort, and persons who lived through hardships or joys together (i, 173). Popper (1952: i, 175) added that Durkheim did not appreciate that economic theories are concerned with abstract relations between people, like exchange of goods and cooperation in the division of labor.

These comments of Popper are largely off the mark. Durkheim (1893) did publish *De la Division du Travail Social*, pointing out that the division of labor makes people dependent upon others, and invites strikes, and showed in *Le Suicide* (Durkheim 1897) that a country's suicide rate is higher, when its economy overheats – *une crise heureuse*, and an abstract relation to that. What is more, Smith did not only envisage abstract relations. According to *The Wealth of Nations*, concrete relations subvert free markets: when sellers of the same good meet for amusement, they end up plotting a price hike. But it is true, Durkheim explained lower suicide rates by more integration in churches and households, as well in the population at large during times of war: three concrete groups.

Popper's (1952) chapter "The autonomy of sociology" argued against an even stronger form of collectivism. Popper just did not think much of explanations of societal phenomena involving factors like the general will, national spirits and group minds (ii, 91). These collective hypotheses – we agree – are decidedly less testable than hypotheses about group integration. However, with hindsight it was too easy to be down on them. The mathematicians Von Neumann and Morgenstern (1944) showed how interactive situations between rational persons can be analyzed using game theory, which provided arguments why in a prisoner's dilemma two rational persons will not attain the common goal. It is difficult to escape the impression that collectivists serving up hypotheses about the general will and group needs, were getting at similar conditions for mutually detrimental interaction.

The call to practice *methodological* individualism has been taken as differing from the declaration that in the end societies and groups do not exist, but only individuals. The latter is the thesis of *ontological* individualism. By postulating an autonomous sociology, Popper accepted the existence of societies. Now, if sociology is to be autonomous, it is perhaps better not to talk about human environments encompassing institutions, but to accept that corporate actors are real. States are very tangible: they tax households, draft men, jail criminals, and refuse passports and visas. Here a new issue crops up: why should sociology go after *methodological* individualism, when it is accepted that societies and corporate actors exist? Durkheim's quest for properties of groups making up societies, may be quite challenging, while corporate actors may have more resources than natural persons, which gives some priority to testing hypotheses about corporate actors relative to those on natural persons. Durkheim (1895) argued in *Les Règles de la Méthode Sociologique* against the mindset of answering societal questions by immediately bringing in individuals. He accepted that societies consist of human beings, but at any one moment societies also comprise concrete groups and institutions, adding that societies comprise material objects too, like houses and roads.

Philosophers of the social sciences like Popper tend to plug rock-bottom explanations, with persons being the fundamental elements of societies. All the same, at least some explanations of societal factors by other societal factors have turned out quite satisfactory, and no explanation ever reaches the lowest – or deepest –

level. Some scholars now hold that sociology's ultimate explanations should refer to human genes, and one also may argue that questions about human societies, should be sub-questions of questions about animal societies in general – how is it possible that human societies have much larger populations, living in peace, than societies of their close genetic relatives?

8.1.3 Multi-level modelling as a theoretical method

There is another argument against reducing sociology to propositions on motives and opportunities of persons. It is implied by Popper's (1935) *Logik der Forschung*, which argues – in section 70 – the impossibility of deriving macro- from micro-statements, since such a deduction requires assumptions about the frequency of micro-entities at the macro-level. Popper's argument – its example was from physics – implies that hypotheses on human societies cannot be derived from theories on human beings: it also should be known how many persons of this and that type make up societies. A national suicide rate cannot be derived from a statement with the chances for Catholics and Protestants to commit suicide: the ratio of Catholics to Protestants in this society should also be known.

Popper's argument seems unbeatable – but backfires. The auxiliary assumption to derive a guess on societies from an individual conjecture should refer to both societies and individuals. Such postulates are logically possible, which brings out that the collectivism-versus-individualism controversy was marred by a false contrast. In addition, statements referring to entities at the macro *and* micro level, may not only form a bridge between the macro-and micro-level. A higher-level proposition may refer to societies *and* their inhabitants, and so may a lower-level one. What is more, sociology's *questions* themselves may involve both individuals and societies. Sociologists studying changes in the degree of income inequality between the inhabitants of societies like the Netherlands, raise a question about two kinds of units, and when they postulate effects of left-wing parties and trade unions, they invoke two more entities. The questions raised are about persons-within-societies. Ultee (1998) brought individuals back in sociology's questions on societal features. That was an attempt to improve by way of concrete cases upon Homans (1964).

To what extent does sociology feature persons-within-societies and societies-comprising-persons? During the UK polemic about individualism or collectivism as a theoretical method for sociology, the tradition of empirical social research – firmly rooted in the USA – witnessed a hunt for 'contextual' effects. Lazarsfeld and his students searched for contextual effects in data on voting and various opinions for the USA in the 1940s and later. The expedition met a big success, when Butler and Stokes (1969) found a contextual effect on voting for Britain: if the district of a manual worker had a higher percent of manual workers, this person's chances to vote Labor were higher. In addition, Lazarsfeld's pupil Coleman (1966) tested the

proposition that grades of US pupils depend on their race, plus the racial composition of their school class. Race was an individual factor explaining individual grades, composition a contextual property. Van Tubergen, Te Grotenhuis and Ultee (2005) reported that for the Netherlands religious persons from a municipality with a higher percent of religious (rather than non-religious) inhabitants, were less likely to die of suicide.

Lazarsfeld and Menzel (1961) tried to codify the logic of research involving entities from various levels. Coleman, Etzioni, and Porter (1970) had a go too. These sociologists took a closer look at sociology's hypotheses than philosophers of the social sciences – like Nagel (1961), who is up next. For two decades now, multilevel statistical models are the rage, allowing for a proper estimation of individual and contextual effects.

Whereas Aristotelian logic only deals with statements containing a subject, copula and predicate, since 1900 relational logic is around, and it is known how to formalize statements involving several kinds of units. It is a pity that Nagel (1961: 535), when commenting on the collectivism-versus-individualism controversy in sociology held that expressions like 'the current President of the United States' only refer to an individual. This depiction bypasses the shift from old to new logic: the phrase refers to two kinds of units, a person and a country, and a relation between them: presiding over. The possibility of formalizing statements involving various kinds of units, has spread within the tradition of empirical social research through Opp's (1970) *Methodologie der Sozialwissenschaften*. If theoretical methods are to be tagged, our merger of collectivism and institutional individualism may be called multi-level modelling. The method advocated by us differs from the older ones by a frank acknowledgment of the existence of corporate actors, and a rejection of the rule to go immediately from the societal to the individual level, bypassing intermediate groups.

We wind up by addressing a question readers may have raised when we first mentioned institutional individualism as a theoretical method for sociology: what exactly is an institution? It must be said, the collectivism-versus-individualism controversy was unclear on what counts as an institution. Popper's (1963: 125, 133) enumeration in *Conjectures and Refutations* includes governments, grocery shops, insurance companies, police forces, schools, and stock exchanges, but counts states out. However, if a state is not an institution, it is a bundle of institutions.

To improve theoretical methods, it is crucial to distinguish first methods referring to persons only, from those postulating natural plus corporate actors. Then both types of methods should be divided by the properties of these entities. Some are 'inherent' – like human instincts and group needs -, others 'relational' – like the opportunities people have, including the walls states build around or in front of them -, and the groups making up societies. In "The aim of science" from 1957, reprinted in *Objective Knowledge* from 1972, Popper held that knowledge grows by turning seemingly 'inherent' properties into 'relational' properties (1972: 195). He

elucidated this thesis with cases from physics. In sociology, Lazarsfeld forged explanations with 'contextual' properties of persons – relations between them and their environment -, and hinted that apparently 'global' properties of societies – like their coherence – may be recast as 'structural' ones – the links between the groups part of these societies. Maritain (1923) had christened Aristotelian logic a logic of inherence, setting it apart from Frege's and Russell's relational logic from around 1900, and Barth (2018) inserted the distinction between inherent logic and relational logic into interesting hypotheses about ways of thinking as tied to well-known persons and societal movements.

For those interested in definitions, the Dutch sociology textbook by Ultee, Arts and Flap (1992), named a group of persons, who maintain ties not only by face-to-face-interaction, but also through an organization created for that purpose, for brevity's sake an institution. The authors of the present paper now would replace in that long sentence the word organization by the expression corporate actor. But then, Popper was not interested in definitions, nor are we. We just defined institutions by stipulation as well as nominally, and avoided – like Popper – an essentialist definition.

8.1.4 Three scandals and organized distrust

In *The Asymmetric Society* Coleman (1982) argued – by borrowing a term from jurisprudence, and findings from Ronald Burt (1975) – that contemporary technologically more developed societies witness an increase in the number of corporate actors, not only bureaus of the state, but also charitable foundations, other tax-exempt organizations like universities, as well as companies with limited liability. This would make for a shift in the balance between natural persons and corporate actors, toward corporate actors. Adam Winkler's *We the Corporations, How American Businesses Won Their Civil Rights* (2018), contains a chronology of corporate rights, corroborating Coleman's claim about increasing asymmetry in favor of corporate actors.

In contrast to the thesis of Coleman (1982) and Winkler (2018), it may be held that some corporate actors work against other corporate actors, and tilt the balance a bit – or more – towards natural persons. At least, this is how we read research by Korpi and Palme (2003). These Swedish sociologists accounted for differences in various social security benefits between industrialized democratic states by variations in the percent of votes for left-wing parties and membership rates of labor unions. If in these countries left-wing political parties and labor unions were stronger corporate actors, the floor in the distribution of income for the natural persons of this country was placed at a higher level. Perhaps Korpi and Palme sinned against institutional individualism, because they did not analyze 'individual data', only data on states and other corporate actors. But then, so what? The power of some corporations may be countered by other corporations, with natural persons as winners.

The theoretical shift from states to corporate actors hints that Popper's theory that organized distrust lowers cheating and violence, may be right for states and their polities, but wrong for other corporate actors. Recent newspaper headlines assert that scientific periodicals publish papers with fake data, carmakers fiddle with emissions of harmful substances, and priests of the Roman-catholic church have sex with minors, while its hierarchy lets things pass. How is distrust in these corporate actors organized, if anything was organized? Three threads will link our analysis of these scandals.

To begin with, Popper admitted that institutions are only as strong as the persons staffing them. We add that special arrangements strengthen the actions of these persons. Knowledge and money for surveillance are obvious requirements, since it is an old trick of un-politics to regulate without proper enforcement. We will try to find out more in this respect for our three scandals, adding that penalties deter society at large, but noting that complaints against those in charge of corporate actors have been settled in silence out of court. So, we vary upon Hobbes (1651) that covenants without the sword are but words, and do not protect human life, their prima facie goal.

Secondly, we wonder to what extent a theory of Merton (1938) applies. In societies like the USA since the 1930s, people would be exhorted to live up to the ideal to get rich and attain that goal now rather than soon, but they lack the legitimate means to do so, making for fraud and infraction of laws in general. We will determine whether the rank and file of corporate actors invent improper actions if the goals held up by their supervisors vastly surpass older goals.

Thirdly – aiming to go beyond the idea that scandals result from 'misfired two-person games' – we hold that scandals feature several actors. Precisely because reviews by scientific journals involve several anonymous persons, fraudulent scientists will not easily be found out, since reviewers may take their task lightly, counting upon other reviewers, who do the same. Something similar may apply to the division of labor within car companies, and the levels of authority in the Catholic church.

A follow-up on the last point. Whereas the term institutional individualism got accepted in Anglo-Saxon sociology around 1970, in Dutch sociology in the 1980s the label structural individualism stuck. Are there structures which are not institutions? Raub and his pupils deserve credit for the attention paid to the prisoner's dilemma – and general game theory – in later Dutch sociology (Raub and Voss 1986; Raub, Buskens, and Corten 2015). So, does the prisoner's dilemma invoke a structure which is not an institution? We hesitate to say: upon closer inspection, not really. The sheriff holds an office, and tries to catch criminals because he faces re-election, whereas his trick to make suspects confess, is under normal conditions impermissible in a lot of present-day societies. Indeed, the prisoner's dilemma has

been taken as a two-person game, whereas its structure involves at least four actors, the two suspects of a crime, each as natural actors, the sheriff as a corporate actor, and a committee specified by law that determines electoral results.

8.2 Scandals in science

'Is the scientific paper a fraud?' That is the question Medawar (1963), Nobel-prize winner and director of the UK *National institute for medical research*, raised in a BBC lecture, printed in *The Listener*. He gave three answers, depending upon how the question is taken. Firstly, the question may be understood as about misrepresentation of facts. If so, the answer is 'of course' no. Secondly, if the question is interpreted as about deliberately mistaken interpretations, the answer is no too. Thirdly, the question may be comprehended as about the representation of thinking in a scientific paper, Medawar's actual target. The answer to the question of whether scientific papers are a fraud in this sense is yes – according to Medawar. Editors of learned journals insist – at least in 1963 biology – on the form a) introduction, b) previous work, c) methods, d) results, e) discussion. That order betrays the false idea that science starts with observation, with theories coming later. Medawar then pays tribute to Popper: scientists should look out for refutations of conjectures.

This section deals with Medawar's first question, and asks how it came about that Diederik Stapel from the Netherlands in the first decade of the 21st century published a series of articles with fake data in refereed social psychology journals. Of course, it may well be that scientists like Medawar would never call (social) psychology a science. Stapel made the *New York Times* (*NYT*) on November 2, 2011, with Carey's "Fraud Case seen as a Red Flag for Psychological Research", and on April 23, 2013, with Bhattacharjee's "The Mind of a Con Man". In 2012, the Dutch science reporter Van Kolfschooten discussed the Stapel case in *Ontspoorde Wetenschappen (Derailed Sciences)*.

A fourth reading of Medawar's question – missed by him – is of interest for sociology. Hume (1757) was too fast in specifying the question of the origin of religion into the question of the origin of religion in human nature, and Durkheim (1912) sought the origin of religion in human societies. So, when it comes to scientific papers, do they twist the social processes by which they come about? Of course, papers are not written with the goal of a correct representation of social processes, but their aim is neither to tell how things happen in the minds of their authors. All the same, because editors prefer papers applying a particular plan, they all follow one format. The balance between the author as a natural person, and the academic periodical as a corporate actor, tilts towards the latter, and makes the question of how Stapel got his papers published more pressing. Stapel for umpteen publications

dreamed up answers to questionnaires, as well as observations in various public places, which may be taken as indicative of, and caused by, a larger mental disturbance, as yet to be identified. However, how did a long list of Stapel's papers pass the roster of referees for a succession of social-psychology periodicals? Why was organized distrust – in the form of reports by referees, who did not learn from a journal's editors the name of a manuscript's author, with that author not learning from the editors the names of those reviewers -, why was this double-blind referee system not effective in keeping out from these journals papers with faked data? Stapel lost his chair at Tilburg University in the Netherlands.

8.2.1 Three older frauds in academia

Since quality newspapers nowadays employ science reporters, irregularities in academia receive publicity, and when Stapel's data faking got out, journalists adduced precedents. We here detail three underplayed scandals. The first one shows that there is more to claims to priority in science than honor, and the second one is about an earlier case in psychology of forged data. The third one is directly relevant to Stapel's case: just as Stapel's students held that his data were 'too good to be true', so the statistician Fisher once argued that the biologist Mendel 'cooked' results.

First the discovery in the 1980s of the retrovirus responsible for acquired immune deficiency syndrome (AIDS), a name coined in the USA in July 1982. We follow findings reported in France (2016). The scientists involved are Robert Gallo from the *National cancer institute (NCI)* in Bethesda, Maryland, USA, and Luc Montagnier of the *Institut Pasteur* in Paris, France. On May 20, 1983 in separate papers in *Science* (France 2016: 104), Gallo and Montagnier described a retrovirus causing AIDS, to be called HIV in 1986. Gallo, the mind driving the field of retroviruses, was competitive and claimed the discovery (p. 216), but had found the retrovirus in a sample from Montagnier, who was branded by Gallo as 'prone to error' (p. 102). In February 1985 science journalists at a gathering in New York got it: here was a case of laboratory intrigue, and low-down thievery (p. 172). The meeting had been called by Mathilde Krim from the *American Medical Foundation*, a privately funded research campaign against AIDS. The *Institut Pasteur* started a law suit in August 1985, which became cantankerous (France 2016: 225). It concerned the royalties for a patent on the HIV-test filed by Gallo and NCI, six months after Montagnier and the *Institut Pasteur* had done so (p. 141). In June 1986 Salk, who was the first to develop a vaccine against polio, and had been embroiled in a similar dispute in the 1950s, mediated and failed. In March 1987 US President Reagan and Prime minister Chirac of France declared – at a joint appearance in the White House – that royalties were to be shared equally, with Gallo and Montagnier named as co-discoverers. When in 2008 the Nobel prize committee singled out the

discovery of HIV, the prize went to Montagnier and team member Françoise Barré-Sinoussi: Gallo did not share in (p. 227). That was the outcome of what sociologist Merton calls a priority dispute. A *Bible*-like battle of David versus Goliath ended with the wise weakling winning.

The second almost forgotten scandal is the Burt affair, after Cyril Burt (1883–1971), an English psychologist. Burt (1966) had published a correlation of 0.771 for the IQs of 53 identical twins reared apart, and of 0.994 for 95 identical twins reared together. Princeton psychologist Leon Kamin (1974) pointed out in *The Science and Politics of IQ* that Burt published a paper in 1955 with exactly the same correlations for 21 identical twins reared apart, and 83 reared together. That coincidence was too strong. It was not clear either how Burt (1955) found his identical twins reared apart, and the number seemed high. The raw data – if they ever existed – were burned with Burt's papers after his death, in accordance with his will. The journalist Oliver Gillie in October 1976 in the London *Sunday Times* was the first to openly suggest that Burt had faked data. From 1947 to 1968 Burt was assistant editor, sole editor, and again assistant editor of *The British Journal of Mathematical and Statistical Psychology*, and Burt often published in this journal. We did not gather how the double-blind referee system – if it already was in place – was applied to his papers.

The Mendel-paradox comes third. Gregor Johann Mendel (1866) published results of experiments on the inheritance of visible traits of garden pea plants – like flower color, and seed shape -, as obtained by artificially fertilizing plants differing in these traits. The statistician Fisher (1936) reworked Mendel's data, concluding that they were too close to theoretical expectations, and wrote that Mendel cooked data. Since Mendel's papers were not archived, and everybody accepts that the Augustinian monk grew peas in the garden of the Abbey of Saint Thomas in Brno, scholars have argued about what exactly Mendel did. We refer to Franklin et al. (2008), *Ending the Mendel-Fisher Controversy*. Mendel redid calculations when results differed too much, and lacked the statistical sophistication to redo them when they agreed. Nowadays standard procedures take care of outliers. They were not around in Mendel's days, and Fisher contributed to their development.

8.2.2 A Dutch social psychologist found out

Stapel studied psychology, got his PhD from Amsterdam in 1997, became a professor in Groningen in 2000, moving to Tilburg in 2006. In August 2011 three young researchers went to the chair of the department of social psychology, and voiced the suspicion that their professor fiddled. The results obtained with data Stapel had collected when – say – visiting schools, were 'just too good to be true'.

When Stapel was confronted by university authorities, he gave in. He had published in *Science* on his observations in the hall of *Utrecht Centraal*, the largest Dutch railway station. According to the faked findings, in waiting areas with litter,

white and black persons were less likely to sit next to each other than in clean areas. By visiting Utrecht, Stapel saw that the location foreclosed such findings, and it dawned on him that he would be found out. A militant vegetarian social psychology professor from Nijmegen, joined up with Stapel for a fake-data paper – in the end unpublished -, showing that people who eat meat are more likely to act selfishly. Suspicions were so concrete, that inquiries were started at every university where Stapel held a job: the *Levelt Committee* for Tilburg, *the Noort Committee* for Groningen, and the *Drenth Committee* for Amsterdam. They followed the same schedule: where are the data for each publication now, what did the questionnaire look like, what were Stapel's hypotheses, and were emails exchanged? Of the 130 articles in Dutch or English, 55 drew on faked data. Of 24 chapters in books to which Stapel contributed, 10 featured fraudulent figures. Another 10 articles probably resulted from deceit too. It is not clear whether Stapel initially altered scores in datasets only, and later invented full datasets, or whether Stapel from early on faked collecting data.

It is easy to criticize the final report of the *Levelt-Noort-Drenth Committees* from November 28, 2012. However, manuscripts are reviewed, academic periodicals keep author names away from referees, and do not give referee names to authors. So, given the import attached in this report to 'double-blind reviewing', an in-depth study of reviews on Stapel's rejects(-and-resubmits) might have been telling. Unfortunately, the committees performed no such analysis.

8.2.3 How did it come about that a social psychologist cooked data for such a long period?

When entering into the question of how Stapel's fraud was possible, newspapers mentioned the pressure to publish, and turned Stapel into the tip of an iceberg. That massive mountain would also comprise phenomena like spreading the content of one big article over several smaller ones with a lot of overlap. However, Stapel had tenure, and was rather relaxed. So, who was forcing him?

Of course, strong organizational demands may compromise objectivity. All the same, it remains to be seen whether some coercion to publish is that bad. Taxpayers deserve value for money, and what presses tenured professors to perform? Blau (1973) studied productivity differences between US university departments. These not only resulted from stronger or weaker selection among applicants. When the climate in a department is more oriented toward research – indicated by the number of graduate students and the import of publications for appointment -, the productivity of a member of this department – measured by the number of publications over a period – is higher. This contextual effect unearthed by Blau may be projected behind the research schools nourished by the Dutch Minister of education since the 1980s. The

Levelt-Noort-Drenth committees bypassed the question of whether Stapel's faked data would have come to light earlier within an active research school.

Merton (1973) assembled articles into *The Sociology of Science*. The ethos of science covers several values, and originality conflicts with disinterestedness. This would make for deviance, like fraud and plagiary (Merton 1973: 309–321). Scientists also may abandon research, withdrawing in teaching or administration. By dissecting case histories, Merton showed that priority disputes do not occur simply because of big ego's, but mainly when bystanders take care that esteem for originality goes to the deserving. New forms of deviance may arise: the France-USA HIV-fight was also about royalties.

Merton (1973: 321) adds that any tendency toward fraud in science is curbed by another of its values: the pursuit of truth. Here he bypasses an important aim of science. Truth is, in Kant's phrase, a regulative idea: after tests with negative results, theories will get *thrown out*. Yet, by which regulative idea are they *ushered in* as worthy of testing? It is content, as discovered by Popper (1935) in *Logik der Forschung*. (A seemingly silly but surely serious elucidation: Dutch weather forecasts for the summer, like 'it may rain tomorrow, but it may be dry too' are definitely true, but without content; the statement 'it will rain tomorrow' says at least something, and the statement 'it will rain tomorrow and the day after' conveys more information.) Merton (1957) offered a content-rich theory: he generalized Durkheim's (1897) anomie theory for suicide to predict high US crime rates, and with a generalization of that generalization Merton (1973) explained deviance in science. Such multilayered and informative theories are rare in sociology.

We return from truth *and* content as regulative ideas in a science, to Stapel's sins. He succumbed to the temptation of presenting upbeat tests of theories from social psychology. So, *truth* was out as a regulative idea. However, did Stapel also commit crimes against *content* as a regulative idea? The answer to this question is a definite no. When Bhattacharjee (2013) interviewed Stapel, Stapel insisted that he loved social psychology – in our reading its grand questions and its quite informative theories – but had been frustrated by its messy experimental data. In his autobiography Stapel (2012) mentions as his favorite theories in social psychology – among others – the theory of cognitive dissonance, the impersonal impact hypothesis, the Laurel and Hardy principle, and terror management theory. These theories are rich in content, and it may well be that Stapel sinned against truth as a regulative idea, because of the many informative theories he not simply found in, but articulated for social psychology. Perhaps he liked their high content so much, that he could not stand their refutation. We leave it to others to counter this conjecture.

8.2.4 The failure of double-blind reviewing as a form of organized distrust in social psychology

Why Stapel faked data is one question, why his faked papers were not stopped another. Of course, the double-blind referee system is not meant to keep out faked-data papers, it is supposed to improve their quality. However, quite a few very bad papers float down the hierarchy of journals, and vetting them is time consuming. In addition, incentives for referees to be constructive are weak: they do not receive a financial reward, and their reputation will not soar, since reports are anonymous – although some journals nowadays list referees once a year. Yet, authors are expected to be honest, and referees are able to spot dishonesties: in claims to a new research question, when presenting hypotheses from the literature, and as a part of data collection and analysis. But such deficiencies of the double-blind referee system, just do not explain how Stapel got a series of papers with faked data out of his office and into academic journals.

The historian of science Thomas Kuhn (1961) has argued that physics textbooks mislead when they present two columns of figures, one with expected results, and one with experimental findings, accompanied by the comment that the figures show reasonable agreement. According to Kuhn, no criterion to determine the degree of agreement is given, nor a justification for the cut-off point. But students – new to the scientific community – take home from the rows – made by reputable scientists – a benchmark for saying that their own results agree well or less well. Kuhn (1961) adds that physicists take a student lab report with overly close agreement as presumptive evidence of data manipulation. So, when Stapel was suspected by his students, their instinct was better than his. Stapel was stopped, but why was he not caught earlier on? His researchers did not nail him right after starting out, since the suspicion that results are too good to be true needs a parade of papers.

Following the blurb of Stapel's (2012) autobiography, Stapel now thinks that if he had acted more cunningly, he regularly would have let research go astray. He did not do so, since he got addicted to ever more beautiful results. Stapel (2012: 242) holds that the *Levelt-Noort-Drenth Committees* reported without speaking with him, and accepted remarks on his personality from various persons in interviews. According to Stapel it was unclear who said what, how interviews were conducted, which questions were asked, whether interviews were taken down by an experienced minutes secretary, and whether interviewed persons agreed with the way remarks appeared in the report.

With these bad jokes, Stapel miscasts the *Levelt-Noort-Drenth Committee*s. They studied all his publications, and it is fair to say that no one ever did that. By combing Stapel's full output, his mode of operation could be traced. Indeed, that way cannot show up in a double-blind review. A referee may have misgivings when reading a published paper of Stapel, but when that person reviews a new paper by

him, this person does not know that it is again a Stapel. Double-blind reviewing does not catch serial offenders, since it hides series. An editor might have suspicions, but these remain qualms, because a serial offender abuses different journals. Although 'the famous French sociologist' Bourdieu filled his own journal *Actes de la Recherche en Sciences Sociales* – from its inception in 1975 until right after his death in 2002 -, we take it that Cyril Burt's decades are over.

So, the answer to the question of how it is possible that Stapel's fakes got through the double-blind referee system, is that they passed precisely because reviews are double-blind. In society at large, when it comes to transgressions of the penal code, the frequency of deviance is supposed to be lower, when the chances to be caught are higher and the punishment is more severe. The corporate actors involved – from police, by way of courts, to prisons – share information with one another, and some people have a longer criminal record than others, while first offenders receive a lower penalty. Scientific periodicals until now – and rather rightly so – do not provide referees with a list of an anonymous author's earlier sins and good works. In addition, social-psychology journals in the years before the Stapel scandal, did not ask authors to make data available for re-analysis.

How now about Popper's (1952: ii, 217) argument that objectivity in science is not ensured by the impartiality of an individual scientist, but forms an outcome of give and take between many scientists? Popper here asserts the import of organized distrust: 'objectivity is closely bound up with the social aspect of scientific method, with the fact that science and scientific objectivity do not (and cannot) result from the attempt of an individual scientist to be 'objective', but from the friendly-hostile cooperation of many scientists.' In the chapter on the institutional theory of progress in *The Poverty of Historicism*, Popper (1957) goes further by arguing that scientific progress is a matter of the free competition of thought, and political institutions that safeguard freedom of thought.

Somewhat in contrast to Popper's hypothesis that progress in an academic field is attained by friendly-hostile cooperation between researchers, we hold that double-blind reviewing is but one form of organized distrust, and a weak one. If manuscripts are presented at scientific congresses first, cons are more likely to be caught. So, double-blind reviewing should be combined with a conference presentation, followed by formal remarks from an assigned referee. Of course, organizing conferences and preparing comments are burdensome, but these activities are more visible than handing in an anonymous review, so more praise for conference organizers and commentators might be forthcoming, if they are mentioned in a published paper. Indeed, departments might refuse money to attend conferences without tough trappings. In addition, since young researchers brought Stapel down, the probability of faked data may be limited by periodic meetings of research schools, in which young researchers thrash work in progress of seniors by re-analyzing their dataset.

Merton held that multiple discoveries are common in science. Our analysis of automobile emissions standards in the second decade of the 20th century will show that they also occur in car research.

8.3 (Un)like *Unsafe at Any Speed*: From the Volkswagen-scandal by way of *Dieselgate* to "Das Kartell"

In 1965 Ralph Nader published *Unsafe at Any Speed. The Designed-in Dangers of the American Automobile*. It was an immediate bestseller, and General Motors signed up private detectives, who tapped the Harvard-educated lawyer's telephone for salacious talk, and hired prostitutes to catch Nader in the act. Politicians picked up Nader's message, leading to air bag and seat belt requirements. In the long run, the number of car-accident deaths dropped dramatically: in the USA in 1965 five deaths occurred for every 100 million miles traveled in cars, and one death for every 100 million miles in 2015. According to our own back-of-the-envelope computations, in 1965 of every 1,000 persons who died in the USA about 26 died in a motor vehicle accident, and 14 in 2015. This decline is not as strong – an increase in the miles travelled in a year partly explains the difference.

On November 25, 2015, *The New York Times* (*NYT*) commemorated that "50 years ago, *Unsafe at Any Speed* Shook the Auto World". The journalist Christopher Jensen quoted Nader as saying that he not only aspired to the level of getting a law through – the *National Traffic and Motor and Vehicle Safety Act* signed by President Lyndon Johnson in September 1966. Nader told Jensen that he also went after a federal agency to implement that law. Why exactly Nader wanted such an agency – nowadays known as the *National Highway Traffic Safety Administration* -, the newspaper does not intimate.

We like to think that Nader cherished two arguments. The first one is a derivation from Hobbes (1651): covenants without the sword are but words, and nowadays weapons house in agencies. The second argument is an analogy, and involves a casual observation: just as the confederate states who lost the American Civil War in 1865 did not pass laws stipulating equal treatment of former slaves – with a century passing before Federal troops escorted black pupils into all-white schools -, so it could not be expected that safety laws accepted at the Federal level, will be implemented by each of the four dozens of states making up the USA a century later. After the murder of President Kennedy, Johnson pushed through various strong-teethed federal laws implementing civil rights.

The *NYT* for its 2015 piece also interviewed a top executive at an automobile company. He liked neither Nader nor his book, but admitted that governments have

a role to play in automotive safety. Regulations level the playing field among automakers, set ground rules where everybody has to do something, and no one has to worry about competitive disadvantages. Safety laws are not only in the interest of car buyers, carmakers on their own cannot limit the dangers of driving.

It so happened that right before the 2015 commemoration, a new cars-scandal erupted. On September 18, the *Environmental Protection Agency (EPA)* of the US Federal government issued a notice of violation accusing *Volkswagen*, the German giant, of installing software to thwart tests for emissions of pollutants. *NYT*-reporter Jack Ewing told in *Faster, Higher, Farther* (2017), the inside story of the Volkswagen-scandal – as its subtitle goes. Dutch journalist Peter Teffer followed with *Dieselgate*, detailing that more German car makers were involved, and pointing at inaction of the European Union (EU) and its 28 member states. Yet Teffer's diagnosis did not go far enough. According to documents presented by Peter Dohmen and Diermar Haranek (2017) in the German weekly *Der Spiegel*, all German carmakers in secret came to agree on the design of diesel engines, producing "Das Kartell". We now review these studies to describe what exactly made for scandals, to determine to what extent they came about by a lack of organized distrust, and to suggest ways of dealing with them. We hunt for informative hypotheses, and sift out pertinent findings.

8.3.1 The Volkswagen-scandal in the USA: Discovering trick devices, with a lawsuit as follow-up

The major sources of energy in pre-industrial societies – whether their inhabitants subsist by hunting and gathering, cultivating gardens, or plowing fields – were human exertion and the harnessing of animals, for instance cows for plowing. Fossil fuels – coal, oil, and natural gas – are the major sources of energy in contemporary societies like the Netherlands. These sources at first made manufacturing the largest employment sector, leading to the tag industrial societies. It remains to be seen whether in future post-industrial societies, where modal persons earn their living in banking, health care, insurance, research, teaching, tourism, and other services, forms of renewable energy will predominate, like hydroelectricity, solar energy, and wind power.

When car ownership spread, it turned out that (post)industrial countries did not have enough oil in their soils – or under their seas -, and that a few other countries exported huge quantities of oil. The latter ones formed a cartel and hiked prices – so much for free markets dominating the contemporary globalized world -, leading to the first world-wide oil crisis of 1973, and the second one of 1979. These calamities in their turn prompted measures to lower oil consumption. Carmakers replaced battleships by compacts, and gasoline by diesel. When it was discovered that diesel engines emit less carbon dioxide – which contributes to global warming and rising sea levels -, and when major European countries began to tax diesel at a

lower rate than gasoline, diesel cars took off. However, they emit nitrogen oxides and fine soot, contributing to smog, asthma, lung cancer, and early deaths. The engineering problem was to cut down on these emissions, and German carmakers claimed in the first decade of the 21st century to have designed diesel motors connected with urea tanks, which scrubbed nitrogen out of the exhaust, and to have installed filters to catch soot. In this way emissions would conform to EU-issued norms. Because of rising euroscepticism, the EU left their enforcement up to each of its member states.

As Ewing (2017) tells the Volkswagen-scandal, the claim of clean diesel rested on laboratory tests ordered by carmakers. That they were suspicious, became public when Volkswagen started to export its diesel cars to the USA – where diesel had not caught on. The company aimed for higher exports of its diesel cars as part of a strategy to become the world's largest carmaker. Because of its corporate structure, with labor unions in its board, and Germany's federal state of Lower Saxony as a major stock owner, it was impossible to cut costs by trimming the number of employees. However, emissions standards were tougher in the USA than in the EU, and fines for violations higher.

In 2013 graduate students from the budget-poor *Center for Alternative Fuels, Engines and Emissions (CAFEE)* of West Virginia University in the USA, conducted research to honor a $70,000 grant. CAFEE had obtained that meager sum from the *International Council of Clean Transportation (ICCT)*, a non-governmental organization that started out as a world-wide network for environmental officials. Its staff comprised former employees of EPA, which hands out penalties for violations of the USA *Clean Air Act*. It contained since 1990 much stronger emission standards, and high penalties.

ICCT knew that Volkswagens had difficulties in meeting (quite low) EU emission standards in tests conducted in European laboratories, and that its makers opposed higher standards because their implementation would be costly. However, they did pass higher US standards in US laboratories, and ICCT initially wanted to check the hunch that Volkswagen installed in diesel cars exported to the USA special parts to meet higher standards, components absent from dirty Volkswagens in Germany. That guess did not bear out: software in US Volkswagens was better in hiding out-of-bound emissions.

The West-Virginia students tested emissions from a Volkswagen Jetta, a Volkswagen Passat, and a BMW, all German passenger diesel cars exported to the USA, and did so in the laboratory, as well as on the road. It may shock experimental sociologists that these students also measured car emissions 'in the real world', but then most people do not drive cars in laboratories. The results for the laboratory tests agreed with those made available by the German car companies. However, when tested on the road, the emissions of two of the three cars turned out to be a multiple of those in the laboratory: the BMW emissions remained the same and acceptable, the Jetta emissions were 15 to 35 times higher than the USA standard,

the exhaust of the Passat 5 to 18 times. The West-Virginia students did not believe their first results, but in the end could not ascribe them to faulty measurements, or other such factors. They wrote up a paper of 117 pages, and presented a summary in March 2014 at a meeting of the *Real World Emissions Workshop*, sponsored by *The Coordinating Research Council*, itself funded by the petroleum industry and major carmakers. In the audience were persons from Volkswagen's emissions compliance department, officials from EPA, and employees of the *California Air Resources Board (CARB)*. Because of the peculiar location of California's major population centers, CARB sets tougher standards than US federal laws. The students only spoke about car A, car B, and car C, but because of the pictures presented, and the high level of knowledge in the audience, it was clear which carmaker violated emissions standards.

Then things started moving. CARB launched a compliance project, and asked Volkswagen about the pollution control equipment in its cars, including how the engine software regulated doses of the urea solution. When the company was not forthcoming, CARB put more staff on the investigation. It gradually dawned on CARB that Volkswagen equipped its diesel engines with a special software device. In laboratories it made for the injection of urea into the emissions, but it was turned off on the road, where the sound of laboratory rollers is absent. When Volkswagen delayed the inquiry by evasive answers, EPA issued on September 18, 2015 a notice of violation against Volkswagen. The Volkswagen-scandal was born, and by early 2017 Volkswagen settled with CARB, EPA, and American car owners for $2,000 million dollars. Lawsuits by car owners against automobile companies may be taken as an outcome of *Unsafe at Any Speed*. In the wake of that book, car owners successfully litigated carmakers for damages in accidents. Five decades later buyers were promised a car with clean diesel, and actually bought one for dirty driving. By the way, diesel not only reduces the health of persons driving a diesel car, but also the well-being of people who do not drive at all.

Ewing's account of the VW-scandal reads like another David-and-Goliath story. Volkswagen is the giant, and loses out because of David-like students. They study at an outpost of academia, struggle on US highways to measure car emissions, and do so with self-made instruments – which often break down. We think a more to the point analogy and more thorough analysis goes beyond two actors. It takes students as Davids too, but brings in two new Goliaths, and devalues Volkswagen. The giants are the agencies handing out stiff penalties. In our close reading of the *Bible*, Volkswagen turns into King Saul: David's master, who – with his varying moods – is not as benevolent and wise as before. In the *Bible* Saul is subdued by David's harp playing, in the Volkswagen scandal it gives Saul headaches.

8.3.2 How did it come about that Volkswagen fiddled?

According to Ewing (2017: 178–179), fiddling will have begun in the early days when software was installed in cars as a stopgap for unexpected difficulties. After that it

developed into a habit, and even later it turned into a competitive cost advantage for Volkswagen, at least before the trick was discovered by regulatory agencies. Supposedly no person at Volkswagen's higher echelons ordered the installation of deceptive software in cars for the US market. We hold it quite possible that upon discovery of tricks at the lower levels of Volkswagen, the higher echelons tacitly agreed with them.

Ewing (2017) also offers a deeper explanation, which reminds of Merton's explanation of high US crime rates. Ewing does not refer to it, and at first postulates a dysfunctional corporate culture – whatever that may be –, and then a succeed-at-all-costs culture, a thesis richer in content. When winding up, Ewing (2017: 261) offers a quite informative hypothesis: 'The pressure to meet corporate goals at any cost is hardly unique to Volkswagen', and 'Practically all corporate scandals stem at least in part from unrealistic targets coupled with draconian consequences for employees who fail to deliver, often combined with outsize rewards for the star performers.' Apart from the clause about big bonuses, this hypothesis is close to Merton's (1938) theory of anomie. The goals held up to persons as legitimate – in the Volkswagen scandal by the chief executive of the corporation that employs them -, cannot be attained with legitimate means, which makes people resort to illegitimate means – cheating with software. Merton (1938) calls this large discrepancy between legitimate goals and legitimate means anomie. Here Merton borrowed from Durkheim. Ewing (2017) adds to Merton's hypothesis that people do not resort on their own to illegitimate means, but avoid being fired upon a failure to reach that goal.

Economic pages of quality newspapers had carried the story that Volkswagen aimed to become the world's largest carmaker by way of diesel. Ewing outlines effects of this strategy in chapter titles: 'By all means necessary' and 'Impossible does not exist'. As a quote from Volkswagen's chief executive attests (Ewing 2017: 62), Volkswagen's goal was to be attained with any means necessary, that is, any means to attain this goal was allowed. And if persons within Volkswagen held that something was technically impossible, the answer was that impossibilities do not exist – according to Ewing a saying around in the company (2017: 90). In came illegitimate means. When a Volkswagen CEO was asked what he would do if engineers insisted that something was impossible, he answered that he would fire them, and bring in a new team, and if that squat were to say that it could not do it, he said he would fire it too (2017: 89). This style of managing emboldened subordinates to behave in the same way toward their inferiors. However, a weakling ran Volkswagen's department seeing to compliance with outside regulations.

8.3.3 Dieselgate discovered or not in Europe: JRC-Ispra and TNO-Delft

After September 18, 2015, the question in Europe was why the Volkswagen scandal had not been outed in Europe. The short answer is that it did come out there, a

short-long answer that other German diesel makers were involved too. The long answer is that neither EU authorities, nor any of its 28 members, made a fuss about the things that became infamous as Dieselgate. The longest answer is that officials washed their hands in innocence and muddy water. Like Pilatus in the *Bible*, the EU claimed that is was outside its mandate to prosecute – carmakers.

This is our summary of Teffer (2017). If there is one good gal or guy in Teffer's account, it is the European Parliament. It decided on December 17, 2015 to appoint a committee to seek testimonies on 'the measurement of emissions *in* the car industry'. It may be nitpicking, but that title is a misnomer. Emissions were measured *by* the car industry itself, and it would have been a case of wise organized distrust to have car emissions measured *by independent agencies*, and to let them decide whether carmakers lived up to the standards specified in emissions laws.

It was known before September 18, 2015 in Europe that European cars surpassed EU limits on emissions of nitrogen oxides. One place was the EU's *Joint Research Centre (JRC)*. In the town of Ispra in Italy, it runs a laboratory for testing automobiles, and it developed PEMS, the portable emissions measurement system. With its help JRC determined car emissions outside the laboratory, and on the road while driving a car. (So, West-Virginia students designed instruments available elsewhere.) According to Teffer (2017: 34), JRC-Ispra reports from 2011 and 2013 show that emissions of various diesel cars on the road were much larger than those in laboratory tests. Teffer (2017: 35) adds that in its 2013 report Ispra suggests that the discrepancies may be accounted for by defeat devices, that made cars perform better in laboratories than on the road. Following Teffer (2017: 75), Ispra told the European Commission at a meeting on November 23, 2010, that on-the-road tests of diesel cars showed emissions 4 or 5 times higher than the norm. *JRC* did not name cars and companies.

That the Volkswagen Passat surpassed emission limits, was known in Delft in the Netherlands, where TNO (*Nederlandse Organisatie voor Toegepast Natuurwetenschappelijk Onderzoek*) has a section for car emissions. TNO – as the name says an organization for applied research in the natural sciences – investigated car emissions for the Dutch *Ministry of Infrastructure and Environment*. The Netherlands has a law that stipulates governmental documents to be made available to anyone who asks for specific ones, and a television program requested emails sent from the Ministry. A mail from January 11, 2012, mentions TNO-tests in the laboratory and on the road for a Passat, with higher values for tests on the road. TNO asked the Dutch company that imports Volkswagen for clarification. There are no minutes for a meeting on April 27, 2012, but an email from TNO to the Ministry from August 27, 2012, shows that Volkswagen did not sent data promised on that occasion. In reply to a question by Teffer, a spokesperson for TNO said – apparently in 2017 – that TNO had decided in 2012 not to pursue its research on high Passat emissions, since it involved conditions deviating from the test for admitting this car on Dutch roads. However, an issue was whether this test was that good.

8.3.4 Why did Dieselgate languish in Europe?

When Teffer insisted, TNO stated that it does not enforce laws. Did TNO press findings on the Dutch ministry doing so? TNO said in 2017 that research for this ministry is not meant to contribute to law enforcement, and that TNO's main goal is to collect data on emissions in the real world. Yet TNO performs applied research, and an ultimate goal of such investigations is to safeguard public health. When Teffer visited JRC in October 2016, its expert on car emissions held that he is no watchdog, and that approval of cars to be driven on the road is a completely different branch of sport: it is the task of national authorities to act on JRC information. But then, why did JRC keep car names classified?

At hearings in 2016 of the special committee of the European Parliament, the in 2010 responsible member of the European Commission said that it was not the goal of the JRC-studies to catch cars and companies in the act, their sole purpose was to determine discrepancies between tests in the laboratory and on the road. But is not it exactly the task of the European Commission to learn from this discrepancy? The European Commission had its own defense: it issues regulations about emissions levels, but EU members were to implement them. Teffer adds: but why was JRC information not shared with national authorities issuing permits for cars to be driven on its roads?

Another justification for the inaction of the European Commission was that carmakers found loopholes in regulations. But then, why did the European Commission leave them open? In addition, the European Commission is there to initiate procedures, not to sit on the chair of judges deciding in a court case. Why did it do so? What is more, the argument of – say – Volkswagen that other carmakers were not persecuted, and all companies were not treated on equal footing, is a chutzpah. No police officer will accept of an individual to be fined for parking where this is forbidden, the argument that owners of other parked cars have not been charged.

The European Commission also might have taken action under its mandate to protect public health. Public health declines when standards are not met, but also when the number of law-abiding cars on the road rises. It may happen that soot and nitrogen oxides emissions by a car fall, but the total amount of soot and nitrogen oxides to which a pedestrian is exposed, goes up: the number of cars explodes, as well as the number of kilometers driven by each car. Emissions laws are to some extent a side show, and *Eurostat* – the EU statistical office – should present figures about causes of death, and chronic ailments. Just as US statistics about the number of death in car accidents per million miles driven may mislead, so may a trend toward tougher standards in emission laws.

All in all, the European Commission – as regards car emissions – did not hit upon the idea that in a budding union, progress occurs by court cases and precedents. Both Ewing and Teffer surmise that the staff of EPA and CARB had a broader view of the goals of their agencies than the staff of JRC and TNO. This is a too easy

explanation. Of course, the EU, unlike the USA, did not have an agency seeing to the implementation of its emissions laws, and this accounts for the discovery of the Volkswagen scandal in the USA. However, that argument amounts to confounding a remedy with a cause. An agency is a remedy, the cause is not only the anomic corporate culture at Volkswagen, but also – as we argue now – fatalism within JRC, TNO and the European Commission. The goals held up by Volkswagen were too high, those set by JRC, TNO and the European Commission too low.

Merton's anomie theory on crime rates improved upon Durkheim's anomie theory for suicide rates. Durkheim (1897) assumed that human desires are unlimited, while human beings lack the means to fulfill them all. According to Durkheim most societies have norms that adjust the goals and means of a person to each other. However, societies with free enterprise and free labor lack such norms, and the goals of persons will surpass their means, leading to suicide, particularly when free markets are booming. Merton (1938) objected to this human-nature explanation: the legitimate goals of the inhabitants of US society since the 1930s vastly outran the legitimate means at the disposal of a lot of them. Merton invokes the American dream, with its from-rags-to-riches stories. The discrepancy between legitimate norms and legitimate means, would make – so holds Merton – for the use of illegitimate means, like various forms of crime.

It so happens that Durkheim (1897: 311) contrasted anomie with fatalism, a situation in which the modest goals that people set themselves are considered illegitimate by society at large, so that 'their future' is 'mercilessly walled in'. He adds that fatalism occurs very infrequently, but it may account for the high suicide rate of married young men, and of married women without children. (Durkheim did not elaborate). When it comes to the effect of fatalism on economic crimes, fatalism in our way of reasoning makes for just doing one's job, like producing parts never to be assembled into cars.

We posit that fatalism prevailed within JRC and TNO. Results of these organizations did not go to the proper places, but these corporate actors just kept up research, perhaps hoping that results one day will be used by authorities issuing standards. The means of JRC and TNO were specific findings on emissions in laboratories and on the road. Their obvious goal was to publicize specific data on dirty emissions, but that was an illegitimate aim. Their legitimate goal was to produce test results for internal reports, which is a humble ambition indeed. This made for fatalism among the rank-and-file.

JRC, TNO, and the European Commission said that emissions surveillance was not their task. Researchers worth their salt, would see to it that their results arrive at the right places, and European politicians up for re-election would go after violation of emissions standards. Few organizations did, their mindset was fatalist: people did not go after modest goals, justified inaction by pointing to a limited task, and no one leaked to a popular car periodical names of cars failing real world tests. European politicians were so estranged from European thought, that they forgot

Hobbes's dictum about covenants without swords. And they did not hit upon the idea that when an authority at a certain level of decision-making issues a law, this authority should provide for an agency at the same level that implements these rules. It is accepted nowadays that when asked to account for one's action, it will not do to state that one was ordered to do so. But apparently, when asked why one did not do anything, it is considered a good excuse that one was not required to do so.

8.3.5 "Das Kartell" discovered in Germany

Ewing (2017) and Teffer (2017) suggest that the problem with nitrogen oxide emissions of diesel cars was, that companies were afraid to tell owners that their cars needed frequent refills of their urea tank, twice as frequent as oil checks. This should not be much of a difficulty for car designers, as this tank may be enlarged. Why the design was not adapted, became clear in Dohmen and Hawranek (2017). They found that other German carmakers came to fiddle with software for their diesels too.

EU laws against cartels are strong, and the EU sees to it that its members implement them. In connection with a case of fixing steel prices, the *Bundeskartelamt* raided on June 23, 2016, offices of Audi, BMW, Daimler, Porsche and Volkswagen. By-products of the catch were documents showing that the companies had made secret agreements about the design of cars with diesel engines. They did so in meetings of their experts. An issue was the size of urea tanks. If they were to become larger, the price of a car would rise too much. But also, if it were too large, the car could not be fitted with a high quality stereo set, and car owners would not be able to transport their golf sticks in their golf bag to their golf court. In the end, the companies decided upon a small tank. Its disadvantage was more frequent refills of urea, and that was reduced by defective software, installed by more companies than the one who was first to be found out. Of course, the decision to go after tanks of the same size, limited competition, and therefore was illegal. When the *Bundeskartelamt* found out, Volkswagen decided to cooperate as chief witness: Volkswagen gave away the other companies, in exchange for a lower penalty. A bit later Daimler confessed, now competing with Volkswagen to become the first chief witness in a cartel case against German automakers. The prisoner's dilemma was there in the Wild West of the USA in the 19th century, and is there now in Germany and the EU.

8.3.6 Car manufacturing and organized distrust

How much organized distrust was there as regards the corporate actor Volkswagen? One might find organized distrust within Volkswagen itself, at the level of Germany, and at that of the EU. In the Volkswagen scandal, organized distrust failed on all three levels.

Did Volkswagen have internal arrangements to check that it lived up to the law? Controlling departments were better staffed in the US than in Europe (Ewing 2017: 109, 125). However, Volkswagen's department – supposed to keep engineers abreast of regulations – lagged behind those of other carmakers. That EPA and CARB strictly enforced pollution rules, was not appreciated enough, and jurists were not in on the exact US penalties for cheating. Volkswagen did not yet have rules in place that would allow a whistle-blower to report a violation without fearing consequences.

In addition, Volkswagen had a board supervising its executive board. Such a structure is common nowadays in many organizations, not only in business enterprises. The main tasks of the supervisory board are appointing new CEO's and selecting outside accountants to check the company's books. For the rest, the ideal supervisory board critically follows the actions of the executive board from some distance. In the case of Volkswagen this did not result in organized distrust. The chief executive of Volkswagen for nearly ten years, Piëch, became president of its supervisory board. In addition, the Piëch family was the main shareholder of Volkswagen. So, Piëch was not independent.

States may influence organizations within the boundaries of their territory. States formulate laws, for example about criteria new cars have to meet to be driven on the roads. They may also create agencies to monitor whether laws are enforced or remain dead letters. As to car emissions, the USA was and is different from Europe. In the USA the federal state acts. The European Commission did not act in the Volkswagen scandal, the right to do so rested with the separate EU member states.

In the USA, 'Washington' wrought 'covenants with a sword'. Since 1990 the Clean Air Act contained strong standards, and high penalties – and apart from EPA, there is CARB. Once irregularities were discovered these agencies bombarded Volkswagen with questions. In contrast, the EU's emissions rules contained loopholes, and had to be enforced by national states, in case of cars more so than as regards other means of transportation. European agencies exist for airlines, the maritime sector, railways, but not motorways (Teffer 2017: 199–206). No state in the EU kept cars off the road after they failed tests. At best member states stipulated minor penalties for makers of polluting vehicles. The most important conclusion attained – after two years! – by the special committee of inquiry of the European Parliament, was that a serious mismanagement had occurred in the measurement of harmful emissions by diesel cars – at the level of the member states as well as that of the EU.

Finally, organizations testing whether cars pass or fail emissions standards, should be independent of car companies. Early on, companies tested their own cars, and it also happened that small firms checked for big producers whether their cars were OK. According to EU-law, a car certified in one country, is allowed on all EU roads. This occasions a race to the bottom. A European agency helps.

8.4 Scandals about sexual abuse of minors by Roman-catholic clerics in various countries

In *Faster*, Ewing quotes an ICCT'er (2017: 175): only if sure, an investigator will accuse a carmaker of cheating on emissions, *since it is like accusing a priest of child sexual abuse* (our italics). We now turn to scandals of sexual abuse of minors by clerics. Part of our argument is that, precisely because people want to be sure before acting, seemingly small-scale scandals fester. Another element is that precisely because cases are not decided in court, but settled behind closed doors, penalties do not deter other clerics. In addition, precisely these arrangements lower the chances that anyone else who was sexually abused as a minor by a cleric, goes to the police, and declares so – or that parents declare that a cleric sexually abused their child. Finally, abuse of minors – although more common than statistics say – is rare. All this makes for a societal process like that involving an emperor without clothes. No one dares to tell him of his nudity, since he is supreme, but once someone has said that the emperor is clad in Adam's attire, the emperor's nakedness is all over the empire.

Sociology's initiators took up the theme of sex and society. Weber's *Gesammelte Aufsätze zur Religionssoziologie* – taken from a journal he co-edited – touch upon the effects of religious ethics on eroticism (Weber 1920–1921: i, 554–556). And Durkheims *L'Année Sociologique* published in 1898 his piece on the topic of incest – in the first issue of his own periodical. By studying Australian aboriginals, Durkheim tried to purge blurring factors. In doing so, he acted not unlike engineers who test car emissions only in laboratories, and sociologists who take experiments as ultimate tests.

Next to Durkheim's question of the causes of a universal incest taboo, stands that of the extent to which people violate it. Frequencies will be small, but are – like suicide rates – societal phenomena: *regularities* occur in the degree to which a society's members *trespass* its *rules*. Durkheim noted incidents of incest, and we discuss his remarks since they throw a peculiar light on Durkheim's thesis on how ideas on superior powers originate. When dissecting an eye-witness report on a corroboree of Australian aboriginals, Durkheim (1912: 309) mentions that at such a staging of a story about primeval ages, sexual rules are broken: men exchange their wives, and incestuous unisons – although in ordinary times condemned as heinous – occur in the open and without punishment. This happens at night, when people are overexcited by music. Perhaps not only Pink Pop and Woodstock are about sex, and drugs, and rock-and-roll. This may account for high corroboree attendance, assemblies that – following Durkheim – bear and nourish the idea of forces surpassing human beings.

In the 1960s countries like the Netherlands witnessed a sexual revolution. The new hormonal anti-conception pill begat more sex before marriage, and while cohabitation without marriage once was frowned upon, it became rife. In several

countries, homosexuality used to be outlawed, and in others publicly ridiculed, nowadays marriages between persons of the same sex are concluded in town halls, with relatives present. It is true, marriages between close kin remain barred. However, as regards the theme of sex and society, more interesting are recent scandals in western countries, showing that Roman-catholic clerics sexually abused juveniles, with Roman-catholic authorities covering up for clerics who sexually molested minors. Reports almost exclusively are about priests abusing boys.

This section mentions relative frequencies of sexual abuse of minors by Catholic clerics, and estimates the percent of clerics who did so, but its main question is why it took so long for persons who were sexually abused, to stand up, and tell that Catholic clerics did so. According to §249 of the Dutch penal code, sexual acts between parents and their minor children are punished by up to six years in prison, as well as sexual acts between a minor and a person who instructs, cares for, or watches over a minor. An earlier version was more concrete, and mentions parents, guardians, teachers and clerics. The abstract formulation implying misuse of authority was chosen, not to leave clerics out, but to encompass other concrete culprits. A contemporary example is sports instructors.

Figures on sexual abuse of minors differ from more familiar statistics on other crimes published by national statistical offices. The latter refer to events during one year, the former state how many persons were abused or not *in the years of their youth taken together*, and how many priests during *their whole career* did or did not abuse a child. This way of counting makes for higher figures, but less so if minors are victimized later on by another cleric, and if perpetrators are serial abusers.

8.4.1 The USA scandal

Sexual abuse by Catholic clergy was first reported in 1985, by US national newspapers. A priest from Louisiana confessed to have molested 11 minors. In 2002, *The Boston Globe* published on abuse of minors in the Boston diocese. This prompted the assembled American bishops to ask *John Jay college of criminal justice* in New York City for a large-scale study. According to *The Nature and Scope of the Problem of Sexual Abuse of Minors by Catholic Priests in the United States* from 2004 (the *Jay Report*), 4% of all US priests from 1950 to 2002 abused at least one minor. All in all, 4,392 priests abused 11,667 minors (Goodstein 2004). Sources were questionnaires to dioceses, urged to search archives.

News on abuse of minors by clerics spread to other countries, and therewith reports were ordered for – among others – Australia, Belgium, Ireland, and the Netherlands. In Australia, 7% of all Catholic priests between 1950 and 2010 were accused of sexually abusing minors (Anonymous 2017). This figure seems higher than the 4% for the USA from 1950 to 2002. We found no comparable figures for perpetrators in the Netherlands, but unique data about victims.

8.4.2 The discovery that minors were abused in the Netherlands

That Dutch Catholic priests sexually abused minors, became a public concern after newspaper *NRC* and broadcaster *Wereldomroep* on February 26, 2010 publicized testimonies of three victims of the brothers and fathers of the *Congregation of Salesians* from one convent plus seminary. This made for 700 new testimonies by persons as minors part of the Catholic educational system. Joep Dohmen's (2010) *Vrome Zondaars* (*Pious Sinners*) contains results of this investigative journalism.

Dutch Catholic education swelled after the 1900 act of compulsory schooling, as well as the 1917 act financing religious schools to the same extent as public ones. Religious pillars were highest around 1960: Catholic children visited Catholic schools with teachers from the Catholic clergy, protestant pupils went to schools with elders as supervisors, while non-religious persons visited public schools. With the 1962 Vatican council the Dutch Catholic pillar started reeling, and it had toppled by 1990. Among Catholic parents, sending children to a boarding school of their own faith, was more common than among Protestant parents (we will give figures). According to Dohmen (2010: 273), in half of the hundred Catholic boarding schools for boys in the Netherlands in 1958, pupils were sexually abused by their clerical teachers or wardens. Orphanages and reformatories, where children and clergy lived under one roof too, were more likely to count victims. Priests abused altar boys, children who came to confession, choir members, scouts, and nephews and nieces. Some cases of abuse lasted years.

Under pressure, Catholic bishops and priors congregations conceded to an independent inquiry into sexual abuse of minors within Catholic organizations. Thus began the *Deetman Committee*, led by a former Secretary of State for Education, publishing in 2010 *Seksueel Misbruik van Minderjarigen in de Rooms-Katholieke Kerk (Sexual Abuse of Minors in the Roman-Catholic Church)*.

Deetman (2010) estimates that between 1945 and 1985 in the Netherlands a few thousand minors were severely sexually abused (meaning penetration) by Catholic clerics. When including less severe cases (like groping), the number rises to 10,000 or 20,000. Misuse decreased at the end of the seventies, when ordination for priesthood had crashed, and troves of clerics left their church.

Compared with studies for other countries, Deetman (2010) contains a novelty: findings for a large sample representative of the Dutch population above the age of 40 years in 2010. (Apparently there was no point in asking younger persons, since their exposure to an educational system run by nuns and priests was too low). The 34,324 respondents with complete data were part of an internet-panel. Of them (Deetman 2010: 564), 31% stated to have been raised as Catholic, and of all these Catholics 9% had been at a boarding school, children's home, orphanage, or seminary. For the 69% not raised as Catholics, this percentage is 4. This confirms what was assumed about Catholic education.

Persons declared whether or not they were, before the age of 18 and against their will, sexually approached by a non-kin adult. Of all Dutch persons older than 40 in 2010, 10% reported as a youth to have experienced unwanted sexual advances by an unrelated grown-up. Deetman (2010: 564) also presents percentages for four subsamples. Of persons raised as Catholic and part of at least one of the four institutions, where juveniles spend day and night, 21% answered yes to the abuse question. The percentage is 22 for persons who have been in at least one of these establishments, but were not brought up as Catholic. If persons were raised as Catholic and did not visit any foundation, the percent is 12, and it is 8 for those who did not grow up as Catholic and never were part of one of the four institutions where young people visit school and sleep.

These data only inform on whether or not a person and potential *victim* was raised as *Catholic*. How many of their *offenders* were active for *Catholic* organizations? That is the key question, and the Deetman (2010: 564) presents pertinent data from a follow-up survey among all 2,454 sexually abused persons. For these abused persons, the percent abused by adults active in Catholic organizations was 6. For people raised as Catholics, once part of an establishment, and sexually abused, 33% of their abusers was active for a Catholic organization. This percent is 9 for sexually abused persons raised as Catholics and never in a foundation. For the whole sample of 34,324 persons, the percent abused as a minor by an unrelated adult from a Catholic organization was 0.6.

Taken together, figures do not tell well on the Dutch Catholic church. The *difference* between the *8%* of sexual abuse for those raised as Catholic, but outside an institution where they stayed during the night, and the *12%* of abuse for persons brought up as Catholic and part of such an establishment, attests to this. So does the *difference* between the *9%* abused as minors by an adult active in a Catholic organization for those raised as Catholic and never in a night-foundation, and the *33%* abused by such an adult, for all abused persons raised as Catholic and part of such an institute.

Figures do not tell well either on growing up in any day-and-night institution. Of minors raised as Catholic, and part of such an institution, *21%* was sexually abused, *similar* to the *22%* for those not raised as Catholic, and having lived in such a place. A follow-up ordered by two Dutch ministries on sexual abuse in institutions for juveniles – as well as on corporeal violence and mental violence – was issued in 2019. It did not involve enough cases for statistical analysis, and is skipped here.

Deetman (2010) advised the creation of an agency to which people could report by hotline to have been sexually abused as minors by clerics, and lodge a complaint, which was to be heard, and might lead to compensation. This *Reporting Centre* issued annual reports, listing the administrative course of calls. To assess the implementation of organized distrust, we quote from page 118 of the English version of its *Report on Activities 2011–2018* (Reporting Centre 2018), its final report.

From spring 2010 – the agency's start – to December 1, 2017 the hotline received 3,712 calls. Of these rings, 1,650 remained a verbal report, 10 a verbal complaint, and 2,056 resulted in a written complaint (a footnote elucidates the discrepancy of 4). Of all written complaints, 251 were withdrawn, and 334 were settled before a hearing – meaning damages were awarded behind closed doors. A board heard 1,471 complaints (now figures square). Of those complaints, 13 more were settled during the hearing, 20 fell outside the board's jurisdiction, 113 were judged inadmissible, 318 without good reason, and 1,002 valid (when summing, we miss 5 cases). At the time of the final report, for 946 valid complaints damages had been fixed, with another 7 complaints having been settled. The difference between 1,002 and 953 (946 + 7) is not accounted for. (Did in these cases people forgo compensation, or was the final report finished before the agency closed?) The 946 valid complaints for which damages had been fixed, were awarded in total 28 million euros. In an NRC news article, Dohmen (2017) quotes that amount from the *Final Report*, adding that – according to documents that are not public – for 403 settlements 13 million euros were paid out, that is, for deals behind closed doors. Reporting Centre (2018: 118) gives 354 (334 + 13 + 7) settlements, and we counted (Reporting Centre 2018: 159–159) a higher number, to wit 402 settlements.

The upshot is that, just as for commercial companies organized distrust takes on the form of professional bookkeeping and independent auditing, so for corporate actors like churches organized distrust involves consistent administration – on top of financial disclosure. These types of organized distrust were not fully implemented in the Netherlands in the 2010s in regards sexual abuse of minors by members of the Catholic clergy during the decades after World War 2.

8.4.3 How did it come about that abuse of minors became public only when minors had grown up?

If punishments are more severe and apprehension chances higher, a state's laws will be trespassed less frequently. This rich-in-content, well-corroborated hypothesis from classical criminology (Beccaria 1764, Bentham 1789) pertains to apprehension and penalties by the state. But how would a state's police learn that a cleric sexually abused a minor? That cleric will have told this child to remain silent, and if a minor brings itself to telling its parents, these parents first go to a local church authority. This power-that-be will check with the cleric named, who may deny. The chances of a police report about the priest are small, since physical traces – if any – will have gone. The cleric's chances of transfer will be high, as well as this cleric's chances of backsliding. If a cleric who abused a minor sexually, is not brought before a public court, sentencing does not deter other priests either.

We now generalize an interesting hypothesis from Jim Rutenberg's piece in the *NYT* for October 22, 2017, on media moguls and prospective actresses, as occasioned

by #MeToo. A major reason for the delay in US movie stars stepping forward, would have to do with out-of-court settlements, in which victimized women obtain financial compensation, and vow silence. In the same way, deals behind closed doors about minors sexually abused by clerics, decrease the chances that parents of abused minors declare #WeToo. The same goes for grown-ups abused as minors. To repeat, the Dutch hotline led to 2,056 written complaints, of which at least 354(=17%) were settled off the record. Before 2010, a couple of well-publicized examples would have given abused persons courage to speak out, but such instances were rare. The balance between natural and corporate actors favors the Catholic church. For the opposite reason, when a few others have already stepped forward, more adults will come forward a bit later to tell that clerics abused them. In addition, just as parents are unlikely to report to the police their own child for misconduct, so the Catholic hierarchy with low probability will tell on its clerics. The prolonged contestation of the separation of Church and State by the Catholic church – in place for the Netherlands since Napoleonic times – will have contributed to the infrequent referral of sexual abuse of minors to the police.

8.4.4 Abuse of minors and organized distrust

Since research is scarce, words were wasted on the question of whether the pledge of celibacy for Catholic clerics attracts gays and lesbians, the question of the extent to which Catholic rites occasion homosexuality with minors, as well as the question of whether Catholic schools tend to do so. We are grateful to the *Los Angeles Times* for collecting questionnaires completed by 1,854 priests in the USA in 2002, and to Andrew Greeley, who analyzed them in *Priests. A Calling in Crisis* (2004).

According to priest and sociologist Greeley, American priests do not differ from ordinary men in several relevant aspects. One idea of conservative Catholics is that abuse of minors entered the Catholic church, because of a growing relative number of homosexual priests. There is some truth in it. Since 5% of all men is homosexual, the 16% homosexuals reported by Greeley (2004:39), amounts to a clear-cut overrepresentation, and speaks against same-sex sex occasioned by their vow not to marry a woman. Indeed, the percent of heterosexual priests living in celibacy was higher than the percent of homosexual priests doing so: 86 versus 63 (Greeley 2004: 41).

Ross Douthat in the *NYT* of March 28 and 30 (2010) harnessed *Jay report* data. Numbers for accused priests and abused minors spike around 1970. The rise was strongest in sexual misconduct over a long period for priests with teenage boys, with short-term misconduct involving children aged 12 years or younger increasing less. Douthat takes this to confirm the thesis that sexual misconduct within the Catholic church climbed because of the sexual revolution in society at large.

Of all imaginable causes of sexual abuse of minors by priests, the media invoked lack of organized distrust the least. How little was there as regards sexual abuse of minors by Catholic priests? The laws of the Dutch state against sexual abuse of minors were in place, and had an eye for misuse of minors by authorities. Yet despite a *School inspectorate*, the state probably monitored Catholic organizations less closely than public ones, and Catholic ones would not have accepted detailed reporting. Catholic congregations ran schools, and what they did outside school hours was out of reach. The Dutch state let the autonomy of church bodies prevail above the well-being of minors entrusted to them. In this way, a law by and large remained words. Supervision was warranted – precisely because Catholics clerics live outside the world. Nowadays the idea that every school should have an in-house child psychologist reporting regularly to the school inspectorate would meet little resistance. It would have been way out – even for boarding schools only – in the 1950s.

Organized distrust was lacking within Catholic organizations too. Culprits did not have to leave their congregation – indeed, that would have been a very severe punishment. (It has occurred, even with cardinals, see Elizabeth Dias and Jason Horowitz 2019). Offenders were reassigned – which may have been something of a reward. Seemingly things were in good order within the Catholic church. It had its own laws, and there was a church court. However, church law contained the directive that in case of misuse by a priest, the victim and witnesses should remain silent.

The Vatican regarded enforcement of church laws as a task for country cardinals: the highest level of a corporate actor delegates duties to a lower level. *Statistics Netherlands* records the number of persons convicted for §249 of the Dutch penal law – and Drukker's (1937) criminal-sociological study on sexual crimes in the Netherlands has a time series from 1911 to 1930. In contrast, the *Statistical Yearbook* of the Vatican publishes the numbers of priests called, but not the number of infractions of particular church laws. Recently the Vatican did some things. Abuse of minors by priests now always should be reported to the police. This amounts to admitting that it often was not done. The limitation in church law for sexual crimes was extended from 10 to 20 years after the event.

8.5 Was organized distrust always there, and was organized distrust always effective?

The present paper replaced a Platonian by a Popperian problem: in the field of politics and society, the misconstrued question of who should rule, should make way for the questions of how bad rulers may be eliminated without bloodshed, and how the damage they do may be minimalized. We did not enumerate in advance forms of organized distrust, and added types while dissecting three scandals.

This contribution's main thesis was Popperian too, and held that organized distrust limits the harm rulers do to the ruled. We replaced Popper's distinction between instinctual and institutional individualism, by that between theoretical methods which postulate individuals only, and methods that bring in natural and corporate actors. The latter of these methods we dubbed, somewhat lightheartedly, multi-level modelling. To assess our central hypothesis, we employed data from investigative journalism. We did so in the understanding that in sociology not so much findings on societies are in dispute, but hypotheses. In everyday life motivational accounts get higher billing than situational explanations, and societal research may show that this should be the other way around. We weighed our main thesis by dissecting three recent scandals: fraud in an academic field, fiddling with software by cars companies, and sexual abuse of minors by clerics of the Catholic church.

8.5.1 Findings on three scandals – outside politics – compared

We were pleasantly surprised with the host of data on our three scandals uncovered by investigative journalism. A difficulty was that data refer to still continuing series of events.

The Stapel scandal in social psychology was closed, but whether Dutch universities took effective measures, remains to be seen. Also, in the wake of more fraud, US social psychology is cleansing the house: flag-bearing experiments – like the Robbers cave experiment by Sherif (1961), the experiments on obedience to authority by Milgram (1963, 1974), the Stanford prison experiment by Zimbardo (1973) – have not been successfully replicated (Carey 2018). As to future frauds, we await a Stapel-like scandal in Dutch sociology. It may not involve fake data, but a massive massage of a mis-constructed multi-moment multi-actor multi-context mega-dataset.

To prop our Stapel dissection, we cited older controversies, but one of them was not closed either. Behind the priority dispute about HIV – settled by the 2008 Noble prize – lurk questions about how AIDS spread. One of them is the extent to which the medical establishment stumbled, public health authorities failed, and *Act Up* – a gay grass roots movement – accelerated the introduction of drugs suppressing the HIV virus. This issue is brought out in the 1987 bestseller *And the Band Played On* by *San Francisco Chronicle* reporter Randy Shilts, Larry Kramer's 1992 play *The Destiny of Me* – situated in New York -, and Robin Campillo's 2017 movie *120 battements par minute* – happening in Paris.

As the main culprit of our second scandal, Volkswagen is in the dock now – with the possibility that a CEO confesses to knowing of trick software -, and the EU designs tough laws, and adjoined agencies.

The abuse of minors by Catholic clerics – the third scandal we dissected – is also on-going. It shifted from the USA to Europe, is reaching Latin America, and

now embraces the complicity of the Catholic hierarchy. Of late it became public that male clerics violated female ones (Horowitz and Dias 2019). All the same, the media tended to bring the scandal down to priests abusing minor males.

To counter this penchant, a new Deetman committee published *Seksueel Misbruik en Geweld tegen Meisjes in de Rooms-Katholieke Kerk (Sexual Abuse of and Violence against Girls in the Roman-Catholic Church)* in 2013. It addresses the follow-up question of sexual abuse of female minors by clerics, but does not break down tables from the *First Deetman Report* (2010) after gender. From some figures (Deetman 2013: 46), and others in Deetman (2010: 52), we estimate that of all Dutch men above 40 years, 0.8% as a minor was sexually abused by an unrelated adult active in a Catholic organization, and that this percentage for women is 0.2. Because of the small number for abused girls (39 out of 173 abused youths), Deetman (2013) omits percentages – although the total sample is unusually large. (Given the earlier on presented 0.6 percent of men plus women abused as a minor by an unrelated adult active in a Catholic organization, the average of 0.2 and 0.8 weighed for the share of women and men in the total sample, should amount to 0.6. It does not, perhaps of rounding errors.) Sexual abuse by clerics of male relative to female minors remains an interesting topic.

Moreover, Deetman missed a chance to deal with the follow-up of whether priests tend to abuse boys, and nuns girls. It so happens that the Reporting Centre's *Report on Activities 2011–2018* has a data-matrix for 860 cases of sexual abuse of minors by clerics leading to compensation, and it yields frequencies for gender combinations. We do not resist temptation, and compute that 3% (24) of all cases involved female clerics. In 17 of them a female minor fell victim to a female cleric, with 7 female clerics abusing a male minor (since these numbers are small, we omit percentages). From the 836 cases with male clerical perpetrators (97% of all cases), 83% (691 cases) pertains to abuse of boys, and 17% (145 cases) of girls. The answer to our same- or other-sex follow-up question is yes.

When explaining scandals, we were guided by three tenets for devising auxiliary assumptions. The first one was to identify, next to regulations, corporate actors seeing to their execution. This idea meant to fill a hole Popper's thesis that institutions should not only be well-designed, but also well-staffed. The police pursues acts forbidden by a country's penal laws, other laws sometimes are put into place by other corporate actors. Whereas an US agency oversaw car emissions, the EU did not have one. As regards abusive priests, the police at most played a small part.

In testing Popper's theory of organized distrust, we also wished to bring in other and older theories, and we mentioned Hobbes's proposition that covenants without swords remain words, and Merton's anomie theory. Hobbes's idea made us hunt for corporate actors who see to it that violations of laws are punished, and Merton's anomie theory was useful for analyzing Stapel's fraud within social psychology, and the Volkswagen scandal: the goals corporate actors like universities and car makers hold up to their employees as legitimate, may be so high that

employees resort to illegitimate means, like faking data and fiddling with software. However, when dissecting the Volkswagen scandal we also returned to Durkheim's complementary theory of fatalism – in governmental agencies overseeing car companies. Since that theory about low legitimate goals set by agencies and ample available means rarely is invoked, we are hesitant about this near-novelty.

Our dissection of the scandal that clerics sexually abused minors, brought in hypotheses from classical criminology. Punishment is designed to deter a person who trespassed the law, from doing so another time. That is specific prevention. However, it is not likely that assigning an abusive cleric to a similar job in another diocese, without it knowing about this cleric's past, will deter this cleric. Supposedly punishment also deters other people from performing the punished act. This is general prevention. In case of priests who sexually abused minors, general prevention was not likely either. The church settled cases behind closed doors, and in exchange for financial compensation, victims and witnesses promised to remain silent. Public punishment deters more than secret sentencing. We also ventured that if punishment is public, people who later become victim of a similar crime – particularly when its frequency is rather low – are more likely to go to the police, and declare that they became a victim of it, compared to dealing with complaints behind closed doors. This may be called the general encouragement effect of public punishment: help for other victims to speak out.

Finally, we had decided not to press our dissection of scandals into the straitjacket of settings with just two actors. Stapel's fraud was discovered, not by academic periodicals, but students. If all of Stapel's papers had appeared in one periodical in a short time, its editor might have noticed that his data were fishy. But Stapel published in many journals, and double-blind reviewing does not bring out that one author always obtains fantastic results. Stapel's students found the results obtained with his data too good to be true. The Volkswagen scandal did not come to light in the US because the US did test, and the EU and its members not. It was known in Europe that Volkswagen's diesels emitted more than allowed. Volkswagen's fiddling with emissions came to light in the USA, because Federal Washington had strict laws, stiff penalties, as well as a special agency, whereas the EU left surveillance of weak standards to its members. Our dissection of sexual abuse involved not only minors and priests, but also parents, church authorities, and the police.

Reviewing the results obtained by our guidelines for the generation of auxiliary assumptions, it may be wondered whether we replaced structural individualism, so prominent in Dutch sociology, by cultural individualism. We do not really think so. Anomie and fatalism indeed are part of a culture enwrapping individuals. But the theoretical section of this contribution, pleaded for statements referring to both corporate and individual actors. In our concrete accounts, corporate actors – like Volkswagen's CEO and directors of research agencies – held up norms to ordinary human beings.

8.5.2 Further questions on various forms of organized distrust – in politics

The general question guiding our dissection of scandals was: if organized distrust is there, and that effective, why three recent scandals? The short answer is that in two scandals organized distrust was absent. It remains to be seen – if the Volkswagen scandal had not come to light in the USA, say because Volkswagen had decided not to export diesels to it -, exactly when it would have become public in Europe. And it may well be that laws on sexual abuse of minors will be trespassed rather frequently as long as the Catholic church stipulates celibacy for its clergy. Laws on sexual abuse of minors by whatever person, are difficult to enforce anyway, since minors do not notify the police.

In one scandal organized distrust was ineffective: double-blind reviewing will not kill off big fraud in academia. We suggested two other forms of organized distrust. The first was periodic meetings of research schools in which junior members flog research in progress of seniors, the second that a paper up for publication, indicates its vetting at a conference. As regards abuse of minors, we proposed a child psychologist in educational institutions. Rules protecting whistleblowers would help employees of corporate actors limiting their legitimate goals, and fostering a fatalist culture. Proper bookkeeping and consistent administration remain important tools for inspection agencies.

To test Popper's theory of organized distrust, we dissected scandals, and it may be wondered why we bypassed a financial scandal. We had decided to spotlight the Netherlands and Holland-within-Europe, and although the 2008 financial crisis hit Holland hard, it did not occur because of Dutch disgraces. Yet, after reading *De Eurocrisis* by Dijsselbloem (2018) – Dutch minister of finance from 2013 to 2017, and chair in those years of the Euro-group of all 17 euro-countries – we almost changed our mind. Dijsselbloem's summary of his observations: 'As always, the effectivity of what we erect together in European context, depends on the power of our institutions, and the extent to which our rulers, including the politicians among them, are like actors master of their lines' (Dijsselbloem 2016: 165, our translation).This quote – without a reference to Popper – is richer in content than Popper's principle that institutions should be well-designed and well-staffed.

Now a pungent criticism of our questions on scandals. Were they attuned to testing Popper's theory of organized distrust? After some pondering, not that closely. Our test involved several assumptions. The central thesis was that organized distrust eliminates or limits damages rulers inflict upon their subjects when making and executing laws. The classical auxiliary assumption going with it, says that periodic general elections are the only or main form of organized distrust in politics. For academic periodicals, car companies, and churches, we pointed to other forms of organized distrust. Another auxiliary assumption held that scandals reveal vast violations of laws and massive personal damages.

A check of the last assumption, made us review our initial decision to skip politics as a test for our central thesis. We still take scandals as proper test cases. However, we bypassed political scandals, as other scandals would test our prime proposition more severely. But then, if no political scandals arise, has organized distrust been omnipotent? The auxiliary assumption of periodic general elections as the only form of organized distrust is *plainly false*, but as to *content* its alternative of more forms of organized distrust is *watery*: these forms lower the damages rulers inflict on their subjects too, but it remains elusive to what extent they do so. So, it is worthwhile to raise questions on the size of the outcomes of these forms of organized distrust: are they as small as implied by the initial auxiliary assumption of one major form of organized distrust? This strategy for testing hypotheses does not start from scandals or other 'negative' effects, but from additional forms of organized distrust as 'positive' causes. As the final number of this show – and food for thought for follow-up questions – we present a recent instance of organized distrust within politics, but emanating outside parliament.

Germany was late in decriminalizing sexual acts between adults of the same sex, and late in giving same-sex couples the right to marry. Was the former the cause of the latter, as the journalist Görlach (2016) implied in "Germany's Retro Record on Gay Rights"? Less than a year later, German parliament gave same-sex couples the right to marry. How did the pertinent proposal make it to the legislative agenda? The left in parliament was too weak, and within the right (CDU-CSU) the more conservative branch (CSU) had the overhand in this matter. Hence a persistent inconsistency between marriage law and a constitutional right to equal treatment.

The law for same-sex couples to marry, landed in German parliament through organized distrust, which may not be recognized quickly as such. Ministers and prime-ministers, in line with organized distrust, answer questions by members of parliament. They also give speeches at public meetings outside parliament, and would flout organized distrust, if they repeatedly ducked questions from the audience. We now posit that a question by an ordinary German citizen to Germany's prime-minister during a public gathering, led to the extension of German laws allowing marriages between a man and a woman, to marriages between persons of the same sex, whether male couples or female couples. Whereas questions from an audience seem less effectual than questions in parliament, in this case an unforeseen question from the audience was surprisingly consequential.

On Monday June 26, 2017 Germany's chancellor Angela Merkel took part in a meeting at the Maxim Gorki Theater in Berlin organized by *Brigitte*, Germany's largest women's magazine. At the end, its readers had an opportunity to ask questions. Few women rose, and a man got up asking Merkel with resoluteness in his voice the simple and short question 'When will I be able to marry my boyfriend, and call him my husband?'. The audience applauded, Merkel rambled, then said she recently had seen how a lesbian couple took care of eight stepchildren, added she opposes same sex-marriage, and finished by saying she would allow a vote in

parliament, with the members of her CDU-CSU party free to vote according to their conscience. On Friday June 30, 2017 that vote was cast. A query from a man in the audience of a meeting organized by a women's magazine caused a stir, and Merkel's reply ended a stand-off in German parliament. We take it, that if someone had addressed this question at a similar meeting a couple of decades earlier to the (then) German chancellor, it would have caused a disturbance in the audience and a scandal in the papers.

The man who addressed Merkel was Ulli Köppe, 28 years young, raised in Saalfeld in the federal state of Thüringen, and earning his living in Berlin as event manager. Did Merkel misspeak, were *Brigitte* readers rightly surprised? According to journalist Lindhout (2017), the German CDU – to avoid a polarizing issue in the 2018 elections – recently discussed a free vote. So, did Köppe's resort to a form of organized distrust advance same-sex marriages in Germany with a few months only?

We here only state this question, but wish to underline its import, and that of similar questions. The question of marriage for same-sex couples is part of a larger question in societal theory, that of the conditions under which laws become stricter or more lenient, and change in other ways. Are more cases around of an effect of organized distrust from outside parliament? As a preliminary: how to pose questions about *The Emergence of Norms*? We allude to Edna Ullmann-Margalit's (1977) book.

To begin with, the question of how norms arise does not refer to any specific norm. This increases the danger of cherry picking when theorizing (Ultee 1996). Auxiliary assumptions are far from ancillary, and here we propose to select *specific* questions before applying *general* theories.

Also, our questions about scandals at first were about the infringement of *substantive* laws, but when analyzing scandals, we began to raise questions on *procedural* laws as causes of transgressions of substantive laws. We compared procedures in one state with those in another state, and in other corporate actors. Sexual acts with minors are forbidden – that is material law -, penalties in one country may be more severe than in another, court cases are conducted behind closed doors or in public, some criminal proceedings are precluded by lapse of time, and more of such adjectives. We also entered into the question of how substantive laws and procedural laws change. So, apart from the division of actors in natural and corporate actors from jurisprudence, we applied its distinction between substantive and procedural law. Indeed, it should be employed in breaking down questions about the emergence of specific norms into more doable smaller questions.

The question of same-sex marriage suggests another rule for picking specific questions on norms. The law against sex between adults of the same sex, held in France until 1791, the Netherlands until 1811, and Germany until 1969. When studying the emergence of a norm at a particular time and place, the auxiliary assumption of no original norm and evolution towards the current norm, may fail: for some time possibly an opposite norm prevails. Questions about the emergence of norms should specify in which period, and from what to what, a change occurred.

Finally, specific questions on norm shifts should invoke unexpected findings. One surprising change in Western countries after World War 2, was that marriage became possible for same-sex couples. Which other remarkable changes are instances of the general question of how norms shift? In passing we pointed to three. Nader (1965) published *Unsafe at Any Speed*, which found strong support in US congress. *Act UP*-Paris speeded up the introduction of drugs against AIDS, and *NRC*-reporter Joep Dohmen (2010) investigated sexual abuse of Dutch minors by clerics, contributing to reparation for victims.

References

Anonymous. 2017. "Schokkende Cijfers over Misbruik door Priesters in Australië." *NRC Handelsblad*. February 07. Retrieved from https://www.nrc.nl/nieuws/2017/02/07/schokkende-cijfers-over-misbruik-door-priesters-australie-6579686-a1544787.

Barth, Else M. 2018. *Empirische Logica: Essays over Logica, Wetenschap en Politieke Cultuur*. Amsterdam: Amsterdam University Press.

Beccaria, Cesare. 1764. *Dei Delitti e delle Pene*. Livorno: Marco Coltellini. In Dutch: Beccaria, Cesare. 2017. *Over Misdaden en Straffen*. Den Haag: Boom.

Bentham, Jeremy. 1789. *An Introduction to the Principles of Morals and Legislation*. London: T. Payne, and Son.

Bhattacharjee, Yudhijit. 2013. "The Mind of a Con Man." *New York Times*, April 23. Retrieved from nytimes.com/2013/04/28/magazine/diederik-stapels-audacious-academic-fraud.html.

Blau, Peter M. 1973. *The Organization of Academic Work*. New York: Wiley.

Burt, Cyril. 1955. "The Evidence for the Concept of Intelligence." *British Journal of Educational Psychology* 25: 158–177.

Burt, Cyril. 1966. "The Genetic Determination of Differences in Intelligence: A Study of Monozygotic Twins Reared together and Apart." *British Journal of Psychology* 57, no. 1-2: 137–153.

Burt, Ronald S. 1975. "Corporate Society: A Time Series Analysis of Network Structure." *Social Science Research* 4, no. 4: 271–328.

Butler, David and David Stokes. 1969. *Political Change in Britain: Forces Shaping Electoral Choice*. New York: St. Martin's Press.

Carey, Benedict. 2010. "Fraud Case Seen as a Red Flag for Psychological Research." *The New York Times*. November 2. Retrieved from https://www.nytimes.com/2011/11/03/health/research/noted-dutch-psychologist-stapel-accused-of-research-fraud.html.

Carey, Benedict. 2018. "Psychology Itself Is Under Scrutiny." *The New York Times*. July 16. Retrieved from https://www.nytimes.com/2018/07/16/health/psychology-studies-stanford-prison.html.

Coleman, James. 1966. *Equality of Educational Opportunity*. Washington: U.S. Government Printing Office.

Coleman, James. S., Amitai Etzioni, and John Porter. 1970. *Macrosociology: Research and Theory*. Boston: Allyn and Bacon.

Coleman, James S. 1982. *The Asymmetric Society*. Syracuse: Syracuse University Press.

Deetman, Wim. 2010. *Seksueel Misbruik van Minderjarigen in de Rooms-Katholieke Kerk. Uitgebreide Versie, Deel 1 en Deel 2*. Amsterdam: Balans.

Deetman, Wim. 2013. *Seksueel Misbruik van en Geweld tegen Meisjes in de Rooms-Katholieke Kerk. Een Vervolgonderzoek*. Amsterdam: Balans.

Diamond, Jared M. 2005. *Collapse: How Societies Choose To Fail Or Succeed*. New York: Viking.

Dias, Elizabeth and Jason Horowitz. 2019. "Pope Defrocks Theodore McCarrick, Ex-Cardinal Accused of Sexual Abuse." *The New York Times*, February 16. Retrieved from https://www.nytimes.com/2019/02/16/us/mccarrick-defrocked-vatican.html.
Dijsselbloem, Jeroen. 2018. *De Eurocrisis. Het Verhaal van Binnenuit*. Amsterdam: Prometheus.
Dohmen, Frank und Dietmar Hawranek. 2017. "Das Kartell. Enthüllt: Die heimliche Absprachen der Autokonzerne." *Der Spiegel*, August 22: 12–19.
Dohmen, Joep. 2010. *Vrome Zondaars. Misbruik in de Rooms-Katholieke Kerk*. Amsterdam: NRC Books.
Dohmen, Joep. 2017. "Kindermisbruik kost Kerk ruim 60 miljoen Euro." *NRC*, December 18. Retrieved from https://www.nrc.nl/nieuws/2017/12/18/kindermisbruik-kost-kerk-ruim-60-miljoen-euro-a1585381.
Douthat, Ross. 2010. "A Time for Contrition." *The New York Times*, March 28. Retrieved from https://www.nytimes.com/2010/03/29/opinion/29douthat.html.
Douthat, Ross. 2010. "The Pattern of Priestly Sex Abuse." *The New York Times*, March 30. Retrieved from https://douthat.blogs.nytimes.com/2010/03/30/the-pattern-of-priestly-sex-abuse/
Drukker, Leonardus. 1937. *De Sexueele Criminaliteit in Nederland, 1911–1930: een Crimineel-Sociologische Studie*. Den Haag: Martinus Nijhoff.
Durkheim, Émile. 1893. *De la Division du Travail Social*. Paris: Alcan.
Durkheim, Émile. 1895. *Les Règles de la Méthode Sociologique*. Paris: Alcan.
Durkheim, Émile. 1897. *Le Suicide, Étude de Sociologie*. Paris: Alcan.
Durkheim, Émile. 1898. "La Prohibition de l'Inceste et ses Origines." *L'Année sociologique* 1:1–70.
Durkheim, Émile. 1912. *Les Formes Élémentaires de la Vie Religieuse*. Paris: Alcan.
Ewing, Jack. 2017. *Faster, Higher, Farther. The Inside Story of the Volkswagen Scandal*. London: Bantam Press.
Fisher, Ronald A. 1936. "Has Mendel's Work Been Rediscovered?" *Annals of Science* 1, no. 2: 115–137.
France, David. 2016. *How to Survive a Plague. The Inside Story of how Citizens and Science Tamed Aids*. New York: Knopf.
Franklin, Allan, A.W.F. Edwards, Daniel J. Fairbanks, Daniel L. Hartl and Teddy Seidenfeld. 2008. *Ending the Mendel-Fisher Controversy*. Pittsburgh: University of Pittsburgh Press.
Gillie, Oliver. 1976. "Crucial Data was Faked by Eminent Psychologist." *Sunday Times* 24: 10–76.
Goodstein, Laurie 2004. "Scandals in the Church: The Overview; Abuse Scandal Has Been Ended, Top Bishop Says." *The New York Times*, February 28. Retrieved from https://nytimes.com/2004/02/28/scandals-in-the-church-the-overview-abuse-scandal-has-been-ended-top-bishop-says.html.
Görlach, Alexander. 2016. "Germany's Retrograde Record on Gay Rights". *The New York Times*, September 26. Retrieved from https://www.nytimes.com/2016/09/28/opinion/germanys-retrograde-record-on-gay-rights.html
Greeley, Andrew M. 2004. *Priests. A Calling in Crisis*. Chicago: University of Chicago Press.
Hobbes, Thomas. 1651. *Leviathan*. London: Andrew Crooke.
Homans, George C. 1964. "Bringing Men Back in." *American Sociological Review* 29: 809–818.
Horowitz, Jason and Elizabeth Dias. 2019. "Pope Acknowledges Nuns were Sexually Abused by Priests and Bishops". *The New York Times*, February 5. Retrieved from https://www.nytimes.com/2019/02/05/world/europe/pope-nuns-sexual-abuse.html.
Hume, David. 1757. *The Natural History of Religion*. London: A. Millar.
Jensen, Christopher. 2015. "50 Years Ago, 'Unsafe at Any Speed' Shook the Auto World." *The New York Times*, November 27. Retrieved from http://www.nytimes.com/2015/11/27/automobiles/50-years-ago-unsafe-at-any-speed-shook-the-auto-world.html.
Kamin, Leon J. 1974. *The Science and Politics of IQ*. New York: Wiley.

Kolfschoten, Frank van. 2012. *Ontspoorde Wetenschap. Over Fraude, Plagiaat en Academische Mores*. Amsterdam: Uitgeverij De Kring.
Korpi, Walter, and Joachim Palme. 2003. "New Politics and Class Politics in the Context of Austerity and Globalization: Welfare State Regress in 18 Countries, 1975–95." *American Political Science Review* 97, no. 3: 425–446.
Kuhn, Thomas S. 1961. "The Function of Measurement in Modern Physical Science" *Isis* 52, no. 2: 161–193.
Lazarsfeld, Paul F. and Herbert Menzel. 1961. "On the Relation between Individual and Collective Properties." Pp. 499–516 in *Complex Organizations. A Sociological Reader*, ed. Amitai Etzioni. New York: Holt, Rinehart and Winston.
Levelt Committee, Noort Committee, Drenth Committee. 2012. *Flawed Science: the Fraudulent Research Practices of Social Psychologist Diederik Stapel*. Tilburg, Groningen, Amsterdam: Tilburg University, Rijksuniversiteit Groningen, Universiteit van Amsterdam.
Lévi-Strauss, Claude. 1949. *Les Structures Élémentaires de la Parenté*. Paris: Presses Universitaires de France.
Lindhout, Sterre. 2017. "Duitse Homo's in Huwelijksroes na Merkels 'Wende'." *De Volkskrant*, July 1, 2017. Retrieved from https://www.volkskrant.nl/nieuws-achtergrond/duitse-homo-s-in-huwelijksroes-na-merkels-wende~bc155eb9/.
Maritain, Jacques. 1923. *Éléments de Philosophie II. L'ordre des Concepts. I Petite Logique (Logique Formelle)*. Paris: Téqui.
Medawar, Peter. 1963. "Is the Scientific Paper a Fraud?" *The Listener 70*: 377–378.
Merton, Robert K. 1938. "Social Structure and Anomie."*American Sociological Review* 3, no. 5: 672–682.
Merton, Robert K. 1957. *Social Theory and Social Structure. Revised and Enlarged Edition*. New York: Free Press.
Merton, Robert K. 1973. *The Sociology of Science. Theoretical and Empirical Investigations*. Chicago: University of Chicago Press.
Mendel, Gregor. 1866. "Versuche über Plflanzenhybriden." *Verhandlungen des naturforschenden Vereines in Brünn*. Bd. IV für das Jahr 1865. Abhandlungen: 3–47.
Milgram, Stanley. 1963. "Behavioral Study of Obedience." *Journal of Abnormal Psychology* 67, no. 4: 371–378.
Milgram, Stanley. 1974. *Obedience to Authority. An Experimental View*. New York: Harper & Row.
Nader, Ralph. 1965. *Unsafe at Any Speed: The Designed-in Dangers of the American Automobile*. New York: Grossman.
Nagel, Ernst. 1961. *The Structure of Science*. New York: Harcourt, Brace & World.
O'Neill, J. (ed.). 1973. *Modes of Individualism and Collectivism*. London: Heinemann.
Opp, Karl-Dieter. 1970. *Methodologie der Sozialwissenschaften. Einführung in Probleme ihrer Theorienbildung*. Reinbek bei Hamburg: Rowohlt.
Pinker, Steven. 2002. *The Blank Slate*. New York: Penguin.
Popper, Karl R. 1935. *Logik der Forschung. Zur Erkenntnistheorie der modernen Naturwissenschaft*. Wien: Springer.
Popper, Karl R. 1952. *The Open Society and Its Enemies. Second Edition (Revised)*. London: Routledge.
Popper, Karl R. 1957. *The Poverty of Historicism*. London: Routledge & Kegan Paul.
Popper, Karl R. 1963. *Conjectures and Refutations: The Growth of Scientific Knowledge*. London: Routledge.
Popper, Karl R. 1972. *Objective Knowledge: An Evolutionary Approach*. Oxford: Oxford University Press.
Rasmussen, Dennis C. 2017. *The Infidel and Friend of Professor Smith*. Princeton, NJ: Princeton University Press.

Raub, Werner, Vincent Buskens, and Rense Corten. 2015. "Social Dilemmas and Cooperation." Pp. 597–626 in *Handbuch Modellbildung und Simulation in den Sozialwissenschaften*, eds. by Norman Braun, and Nicole J. Saam. Wiesbaden: Springer VS.

Raub, Werner und Thomas Voss, 1986: "Die Sozialstruktur der Kooperation rationaler Egoisten." *Zeitschrift für Soziologie* 15: 309–323.

Reporting Centre Sexual Abuse within the Roman Catholic Church in the Netherlands. 2018. *Report on Activities 2011–2018*. Utrecht: Libertas Pascal.

Rutenberg, Jim. 2017. "A Long-Delayed Reckoning on the Cost of Silence on Abuse." *New York Times*, October 22. Retrieved from https://www.nytimes.com/2017/10/22/media/a-long-delayed-reckoning-of-the-cost-of-silence-on-abuse.html.

Sherif, Muzafer. 1961. *The Robbers Cave Experiment: Intergroup Conflict and Cooperation*. Norman, Oklahoma: University of Oklahoma Press.

Smith, Adam. 1776. *The Wealth of Nations*. London: W. Strahan, and T. Cadell.

Stapel, Diederik A. 2012.*Ontsporing*. Amsterdam: Prometheus.

Teffer, Peter. 2017. *Dieselgate. Hoe de Industrie Sjoemelde en Europa Faalde*. Amsterdam: Q.

Topitsch, Ernst. 1958. *Vom Ursprung und Ende der Metaphysik*. Wien: Springer.

Ullmann-Margalit, Edna. 1977. *The Emergence of Norms*. Oxford: Oxford University Press.

Ultee, Wout C. 1996. "Do Rational Choice Approaches Have Problems?" *European Sociological Review* 12, no 2: 167–179.

Ultee, Wout C. 1998. "Bringing Individuals Back into Sociology. Three Aspects of Cohesion in Dutch Society during the 20th Century." Pp. 188–203 in *Rational Choice Theory and Large-Scale Data Analysis*, ed. Hans-Peter Blossfeld and Gerald Prein. New York: Routledge.

Ultee, Wout C., Wil A. Arts, and Henk D. Flap. 1992. *Sociologie: Vragen, Uitspraken, Bevindingen*. Groningen: Wolters-Noordhoff.

Van Tubergen, Frank, Manfred Te Grotenhuis, and Wout Ultee. 2005. "Denomination, Religious Context, and Suicide: Neo-Durkheimian Multilevel Explanations Tested with Individual and Contextual Data." *American Journal of Sociology* 111: 797–823.

Von Neumann, John and Oscar Morgenstern. 1944. *Theory of Games and Economic Behavior*. Princeton, NJ: Princeton University Press.

Weber, Max. 1920–1921. *Gesammelte Aufsätze zur Religionssoziologie*. Tübingen: J.C.B. Mohr.

Winkler, Adam. 2018. *We the Corporations. How American Businesses Won their Civil Rights*. New York: Liveright.

Zimbardo, Philip, Craig Haney, and Curtis Banks. 1973. "Interpersonal Dynamics in a Simulated Prison." *International Journal of Criminology and Penology* 1, no 1: 69–97.

Rainer Hegselmann

9 Polarization and Radicalization in the Bounded Confidence Model: A Computer-Aided Speculation

Abstract: In the bounded confidence model agents update their opinions by averaging over all opinions that are not too far away from their own opinion. The article gives a precise definition of the basic model, offers several interpretations of the model, and introduces two simple extensions that allow to analyze polarization and radicalization. The basic model and its two extensions are seemingly simple. But the simplicity is deceptive. Lots of counterintuitive effects come as a surprise. Additionally, the article demonstrates that in terms of explanatory understanding of mechanisms, it makes a lot of sense to work with deterministic idealizations of random start distributions.

9.1 Introduction

The so-called *bounded confidence model* (BC model) spells out a simple idea: In their ongoing exchange of opinions, individuals (or *agents*) take seriously those others whose opinions are not too far removed from their own. The core of this article are two extensions of the basic BC model. The *first* extension (Section 9.2) introduces a *bias* of the type: "Leftists listen more to the left, rightists listen more to the right." We show how, given the bias, strongly polarized camps of radicalized agents can evolve endogenously. In the *second* extension (Section 9.3) a group of radicals enters the field. Different from normal agents, the radicals simply stick to their radical opinion – no other view is taken seriously. We analyze how normal agents that take seriously opinions that are not to far removed from their own opinion, may or may not become radical.

This article also has a *methodological* objective. I want to demonstrate that the following is a fruitful approach: *First*, we initialise all our opinion dynamics with the same, constant start distribution. It is a very special, namely *representative* start distribution: it idealizes deterministically high numbers of random start distributions. *Second*, in tiny steps, a selected parameter is varied. Then, by direct comparisons of single runs of the system, we try to get an understanding of the dynamics. In contrast to laboratory experiments, such an approach is easy to implement with the help

Note: Parts of this article draw on two articles that I wrote together with Ulrich Krause, namely Hegselmann and Krause (2002, 2015). I want to thank Matthew Braham and Igor Douven for all their help.

Rainer Hegselmann, Frankfurt School of Finance & Management

https://doi.org/10.1515/9783110647495-009

of computers. Our computational analysis of polarization and radicalization in Sections 9.2 and 9.3 applies this approach. In the concluding Section 9.4 we discuss the advantages, limits, and dangers of our strategy.

Finally a warning: What follows is not the presentation of an empirically calibrated model. It is more a *computer-aided speculation* on the consequences of mechanisms that, as some empirical findings suggest, seem to be at work in the real world. I analyze the idealized mechanisms in an artificial world wherein nothing else interferes. The hope is to get thereby a better feeling for what might be at work in the real world. However, there is also bad news: My computer-aided speculations suggest that there might exist a fundamental problem for both the understanding of and intervening in polarization and radicalization processes.

9.1.1 The bounded confidence model: A formal description and some possible interpretations

Stated in a precise language, the constitutive assumptions of the BC model are:

1. There is a set I of n agents; $i, j \in I$.
2. Time is discrete; $t = 0, 1, 2, \ldots$.
3. Each individual starts at $t = 0$ with a certain opinion, given by a real number from the unit interval; $x_i(0) \in [0,1]$.
4. The profile of opinions at time t is $X(t) = x_1(t), \ldots, x_i(t), \ldots, x_j(t), \ldots, x_n(t)$.
5. Each agent i takes into account only 'reasonable' others. Reasonable are those individuals j whose opinions are not too far away, i.e. for which $\left|x_i(t) - x_j(t)\right| \leq \epsilon$, where ϵ is the *confidence level* that determines the size of the *confidence interval*.
6. The set of all others that i takes into account at time t is:

$$I(i, X(t)) = \left\{j \mid \left|x_i(t) - x_j(t)\right| \leq \epsilon\right\}. \tag{9.1}$$

7. The agents update their opinions. The next period's opinion of agent i is the average opinion of all those, which i takes seriously:

$$x_i(t+1) = \frac{1}{\#(I(i, X(t)))} \sum_{j \in I(i, X(t))} x_j(t). \tag{9.2}$$

The description 1. to 7. gives some very general interpretation of the symbols i, j, n, t, I, $x_i(t)$, $X(t)$ which otherwise would be terms in a pure and 'naked' formalism. We get specific interpretations if we think of certain contexts in which the mechanisms described by equations (9.1) and (9.2) might be at work. Some such (partially overlapping) interpretations are the following:

- *Expert interpretation*: There is a group of experts on something. Each expert has an opinion on the topic under discussion, for instance the probability of a certain type of accident. Nobody is totally sure that he/she is totally right. Within bounds everybody is willing to adjust and to adapt: An expert i considers as competent all experts j that are not too far away, i.e. for which $|x_i(t) - x_j(t)| \leq \epsilon$. The updated compromise opinion is the arithmetic mean over all opinions within i's confidence interval. Such opinion revisions produce a new opinion distribution which may lead to further revisions of opinions, and so on.[1]
- *Compromise interpretation:* There is a group of people that exchange views which can reasonably described by real-valued numbers.[2] For reasons as uncertainty, respect for others that seem to be as reasonable as oneself, an interest in a compromise, a preference for conformity, or due to some social pressure, everybody is, at least in principle, willing to compromise with others.[3] However, the willingness to compromise is bounded: An agent i with view $x_i(t)$ is willing to compromise with agents j with view $x_j(t)$ iff j's view is not too far away, i.e. for which $|x_i(t) - x_j(t)| \leq \epsilon$. Then averaging.
- *Social media interpretation:* There is a digital platform with a central algorithmic coordination that brings together with user i those users j whose opinions x_j are not too far away from i's opinion x_i. Then averaging. In such a context ϵ is the 'distance tolerance' of a *centrally* organised *filter bubble* (Pariser 2011). Another variant is a digital platform that allows all users j to send their opinion x_j to all other users i. As a receiver, user i reads only opinions that are not too far away from i's opinion x_i. Then averaging. Under this interpretation, ϵ is the distance tolerance of a *de*centralised *echo chamber* (Sunstein 2017).

This is not a complete list of possible interpretations. We could also use the BC mechanism as a kind of *aggregation devise*, and look for the smallest ϵ that, for a given distribution of start values, leads to a consensus.[4] From a *normative* point of view, we could use the model to analyze the effects of different recommendations

1 Obviously, that is the procedural format of so-called *Delphi studies*.

2 That includes much more than opinions on the probability of any quantitative or qualitative proposition. As long as one can reasonably normalise the range of possible opinions to the unit interval, opinions could regard any real-valued quantitative problem. The opinions could express the intensity or importance of a wish (though only under the condition of intersubjective comparability). The exchange of opinions might be about moral praiseworthiness (0: extremely bad; 0.5: neutral; 1: extremely good). Or the opinions could regard a budget share. *Not* covered are non-continuous opinions (for instance, discrete or binary).

3 The overview article Flache et al. (2017) mentions and shortly describes many empirical findings from different fields that I interpret as supporting evidence.

4 This interpretation gives the BC model a similar status as it was originally claimed for the Lehrer/Wagner model in Lehrer and Wagner (1981).

for how to resolve peer disagreement and other epistemic problems (Douven 2010; Douven and Kelp 2011; Douven and Wenmakers 2017). A *technical*, namely control perspective is taken in Hegselmann et al. (2015).[5]

9.1.2 Basic features of the BC dynamics

For certain ϵ-values the BC dynamics is trivial: If $\epsilon = 0$, then all agents stick to their opinions forever; if $\epsilon = 1$ then all agents' opinion is the arithmetic mean of the start profile $X(0)$ from $t = 1$ onwards. But what's about the region $0 < \epsilon < 1$? Figure 9.1 shows the dynamics for $\epsilon = 0.05,\ 0.20$, and 0.25. The x-axis is the (discrete) time t; the y-axis represents the opinion space $[0, 1]$. The grey graphs show the trajectories of the opinions of 50 agents.

For the analysis of the dynamics two concepts will be very fruitful: First, we call an opinion profile $X(t) = x_1(t), x_2(t), \ldots, x_i(t), \ldots, x_n(t)$ an *ordered profile* iff $0 \leq x_1(t) \leq x_2(t) \ldots \leq x_i(t) \ldots \leq x_n(t)$. Second, we call an ordered profile an *ϵ-profile* iff for all $i = 2, \ldots, n$ it holds that $x_{i+1}(t) - x_i(t) \leq \epsilon$. Thus, in an ϵ-profile neighbouring opinions $x_i(t), x_{i+1}(t)$ mutually influence each other. In Figure 9.1 vertical grey lines between *two neighbouring opinions* indicate that the distance between the two is $\leq \epsilon$. As a consequence, a continuous vertical line from $x_1(t)$ to $x_{50}(t)$ means that $X(t)$ is an ϵ-profile in the sense defined above.

For an explanation of important effects, we start with the dynamics in the middle of Figure 9.1. That is the case $\epsilon = 0.2$. Careful inspection, time step by time step, shows several effects:

- Extreme opinions are under a one-sided influence and move direction center. Therefore, the range of the profile starts to *shrink*.
- At the extremes of the shrinking profile, the opinions *condense*.
- Condensed regions *attract* opinions from less populated areas within their ϵ-reach: In the center opinions > 0.5 start to move upwards, opinions < 0.5 start to move downwards.
- The ϵ-profile *splits* in $t = 5$. From now on the split sub-profiles constitute different 'opinion worlds', i.e. two communities without any influence on each other.

5 In substance, the model was described for the first time in Krause (1997). In a conference presentation in 1998, the naming "bounded confidence" was used for the first time; see Krause (2000). A first comprehensive analytical and computational analysis of the BC model was given in Hegselmann and Krause (2002). For the history of the BC model and a systematic classification of alternative models of opinion dynamics see Sections 1–3 of Hegselmann and Krause (2002). The closest relative of the BC model is the model in Deffuant et al. (2000); for the similarity see Urbig et al. (2008). For overviews on other, later, and future developments see Lorenz (2007), Xia et al. (2011), Sîrbu et al. (2017), Flache et al. (2017).

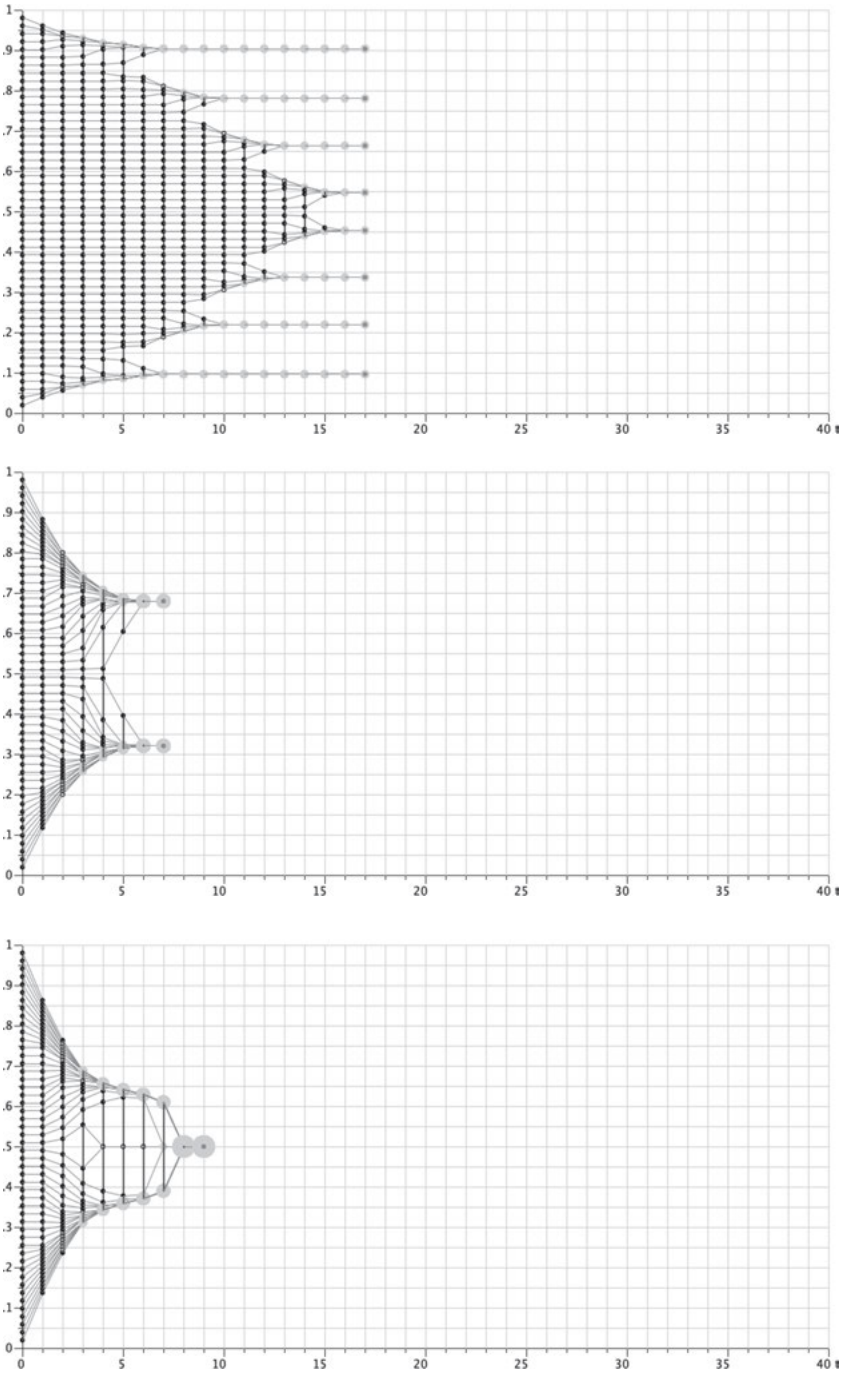


Figure 9.1: BC dynamics for 50 agents.
Notes: Top: $\epsilon = 0.05$. Middle: $\epsilon = 0.2$. Bottom: $\epsilon = 0.25$. Filled black or grey circles indicate the size of a cluster. A black circle is just one agent. A black circle with an inner white circle represents two agents. Grey circles are scaled according to their cluster size.

- In the two split off sub-profiles opinions *contract*. In $t = 5$, in the two sub-profiles all opinions have all opinions within their confidence interval. Therefore, in $t = 6$ in each sub-profile all opinions merge into one and the same opinion.
- As a consequence, we get $X(7) = X(6)$. The dynamics is *stabilized* in the sense that $X(t) = X(t+1)$.

The dynamics in Figure 9.1, middle, ends up with two camps, one to the left, one to the right of the center of the opinion space – a kind of *polarization*. Figure 9.1, top, shows a dynamics based upon the much smaller $\epsilon = 0.05$. The ϵ-profile splits twice in period 6, leaving behind a big ϵ-sub-profile in the center that splits again in period 9. Some more splits occur some periods later. Finally 8 clusters of opinions survive. Let's call such a pattern *plurality*. Figure 9.1, bottom, is based upon a confidence interval $\epsilon = 0.25$. In this case, the ϵ-profile never splits. A cluster of two opinions in the center contracts outer opinions such that, finally, all opinions merge into one – *consensus*. In all cases, the final pattern of fragmentation – plurality, polarization, or consensus – is brought about by shrinking, condensing, attracting, contracting, splits, and mergers. And it is always brought about in *finite* time. For the latter result, simulations are not necessary: for a rigorous proof see Hegselmann and Krause (2002, Theorem 6).

The differences in the three examples in Figure 9.1 are all due to different ϵ-values. The start profile $X(0)$ is *always the same*. It is a profile for which it holds that

$$x_i(0) = \frac{i}{n+1}, \forall i = 1, \ldots, n. \tag{9.3}$$

Such a profile is at the same time very specific *and* representative: The *i*th opinion is exactly there, where it will be at the average over infinitely repeated draws of n opinions that are *uniformly* distributed on the unit interval. Thus, equation (9.3) gives the *expected value* of the *i*th opinion of a uniform distribution of n opinions. We refer to that type of start distribution as the *expected value distribution*. It is a kind of *deterministic idealization* of a certain random distribution.[6] The expected value distribution generates a start profile in which the distances between two neighbouring opinions are always the same: the start profile is *equidistant*.

In all what follows, we will rely on *one and the same* expected value start distribution with $n = 50$. For that constant start distribution we will analyze the effects of, for instance, stepwise increasing ϵ-values. That will be done by careful inspections and comparisons of *single runs*. On purpose, *programmatically*, we deviate from the

6 We can do the same with regard to other types of random distributions. However, one has to derive the equations that then correspond to equation (3). The equidistance is due to the uniform distribution that is deterministically idealized by equation (3). If, for example, we do the same with a normal distribution, the corresponding expected value distribution would *not* be equidistant. The relevant discipline is *order statistics*.

usual practice to run, firstly, a major number of random initialisations, and then, secondly, to do some statistics on the runs. Our approach will blind us to effects that depend largely on the randomness of initialisations.[7] However, our approach may expose directly effects that are otherwise hidden in averages. Anyhow, I postpone the discussion of advantages and disadvantages to the concluding Section 9.4.

9.1.3 A warning: The model's simplicity is deceptive

The three dynamics in Figure 9.1 suggest a more general idea about the confidence level ϵ: The number of finally surviving and stabilized clusters decreases as ϵ increases. A first diagram – we will call it an *ϵ-diagram* (see Figure 9.2) – supports that idea.

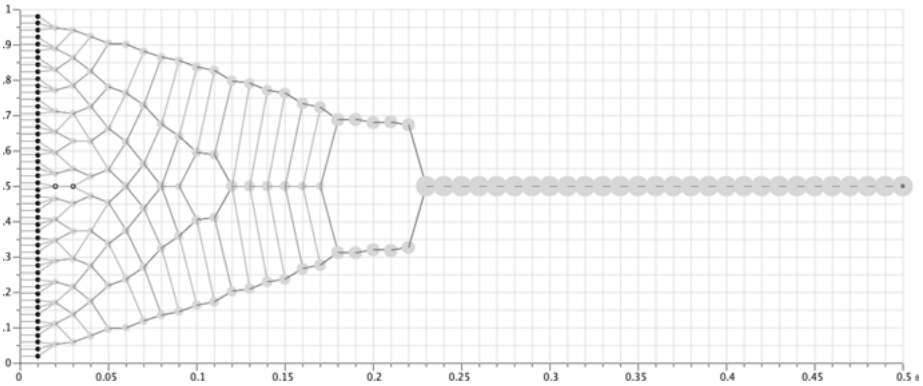


Figure 9.2: ϵ-diagram for an expected value start distribution with $n = 50$.
Notes: On the x-axes ϵ increases with step size 0.01. For a given ϵ, the y-axis indicates the stabilized end positions of opinions. The lines connect ranks in the ordered profiles. Different sizes of grey circles indicate the size of an opinion cluster. A small black circle is a cluster with just one opinion. A small black circle with a white inner circle indicates a cluster of two opinions.

An ϵ-diagram visualizes for *one and the same* start distribution $X(0)$ the effects of a stepwise increasing ϵ on the final, completely stabilized cluster structure. Figure 9.2 is an ϵ-diagram. The x-axis does *not* represent time rather than increasing values of ϵ. For each ϵ value, we run the dynamics until it is stabilized, i.e. $X(t+1) = X(t)$.

7 For instance, whatever n, our expected value start distribution generates *equidistant* start profiles – random start distributions do *not*. But that may have consequences for the number of finally surviving clusters.

The y-axes of an ϵ-diagram is used to display the stabilized positions. Lines connect the positions of ranks in the ordered profiles.

The lessons from the ϵ-diagram in Figure 9.2 seems to be very clear: For confidence intervals in the range $[0,\ 0.17]$ the number of final stabilized clusters monotonically decreases from 50 to 3 – *plurality*; in the range $[0.18,\ 0.22]$ we get *polarization*; for $\epsilon \geq 0.23$ the final result is always *consensus*. Thus, the general idea about the effects of the confidence level *seems* to be: Given the same, constant start distribution, for increasing values of the confidence level, the number of finally stabilized clusters decreases monotonically; we have a monotonic transition from plurality to polarization, and from there to consensus.

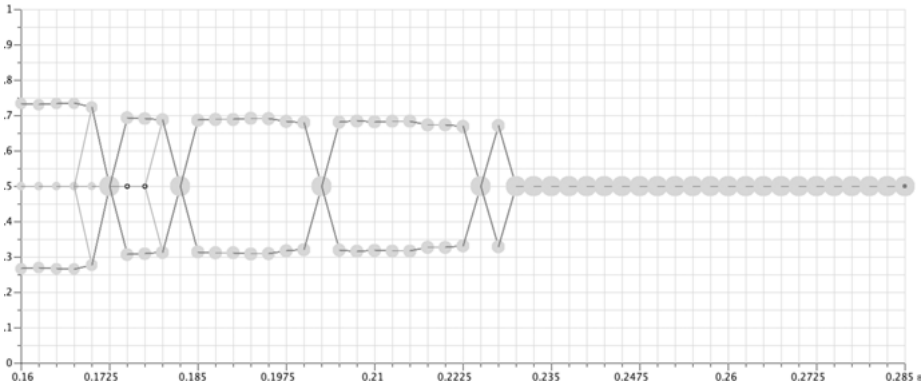


Figure 9.3: ϵ-diagram for an expected value start distribution with $n = 50$.
Notes: Different from Figure 9.2, ϵ increases on the x-axes with step size 0.0025; start is at $\epsilon = 0.16$. For other details see the caption of Figure 9.2.

Intuitive as it may be, the general idea is actually wrong, and Figure 9.3 makes that very clear.[8] As Figure 9.2, Figure 9.3 is an ϵ-diagram for an expected value start distribution with $n = 50$. The only difference is the step size of ϵ: it is now much smaller, namely 0.0025. The smaller step size reveals that the transition from polarization to consensus is not monotonic: A consensus that is reached for a certain ϵ, may fall apart under a greater ϵ. There is always an ϵ^* such that for all $\epsilon \geq \epsilon^*$, consensus is guaranteed.[9] But an actual ϵ that leads to consensus does not necessarily have to be ϵ^*. The non-monotonicity is not confined to the transition from two to one final clusters. And it is not an effect (or artifact) of the equidistant expected

8 Wedin and Hegarty (2015: 2016) write about the model: "The update rule is certainly simple to formulate, though the simplicity is deceptive."
9 This is an *observation*, not a proof. For $\epsilon = 1$ we get trivially a consensus. But, as a matter of fact, normally ϵ^* is much smaller than 1.

value start distribution: the same type of non-monotonicity holds for random start distributions as well.[10]

9.2 Extension one: Polarization by biased confidence

The confidence intervals that we considered so far were symmetric: whatever the total size of the interval, the left and the right part were always of equal size. But what if individuals that hold a more left [more right] opinion 'listen' more into the left [right] direction – and the more to the left [right] their opinions are, the more so. The idea suggests an opinion dependent *asymmetry* of the following type: The more left [right] an opinion is, the more the confidence interval is *biased* direction left [right]. For opinions closer to the center of the opinion space, the bias is less pronounced. Only for opinions exactly in the center the confidence interval is symmetric.

9.2.1 Modelling biased confidence

To model the intuitive idea precisely, we divide a confidence interval of any size into a left and right proportion β_l and β_r, such that $0 \leq \beta_l, \beta_r \leq 1$ and $\beta_l + \beta_r = 1$. Following our intuition stated above, the values of β_l should be given by a monotonically decreasing function $f(x)$, defined over our opinion space $[0,\ 1]$. That, then, allows us to get β_r as $1 - \beta_l$. Since we request symmetry of the confidence interval for $x = 0.5$, it should hold that $\beta_l(0.5) = 0.5$.

One can get all what we want by the simple linear function

$$\beta_l(x) = mx + \frac{1-m}{2}, \tag{9.4}$$

where m, the slope of the function, controls the strength of the bias. A reasonable range for m is $[-1,\ 0]$. Figure 9.4 shows the resulting functions. For all values of m, the function rotates around $\langle 0.5, 0.5 \rangle$. The meaningful functions are between the dashed graphs with the slope $m = -1$ and $m = 0$. For the strongest bias $m = -1$, the most extreme leftist [rightist] opinion, that is $x = 0$ [$x = 1$], the right [left] portion of the confidence interval is 0. For $m = 0$ the opinion dependent bias disappears: For all opinions x it holds that $\beta_l = \beta_r = 0.5$ – the confidence interval is symmetric. The black graph is an example for a bias in between the extreme values for m. Given the black

10 Note that for any type of constant start profile (e.g. random or expected value), the width of an opinion profile does also not decrease monotonically with an increasing confidence level.

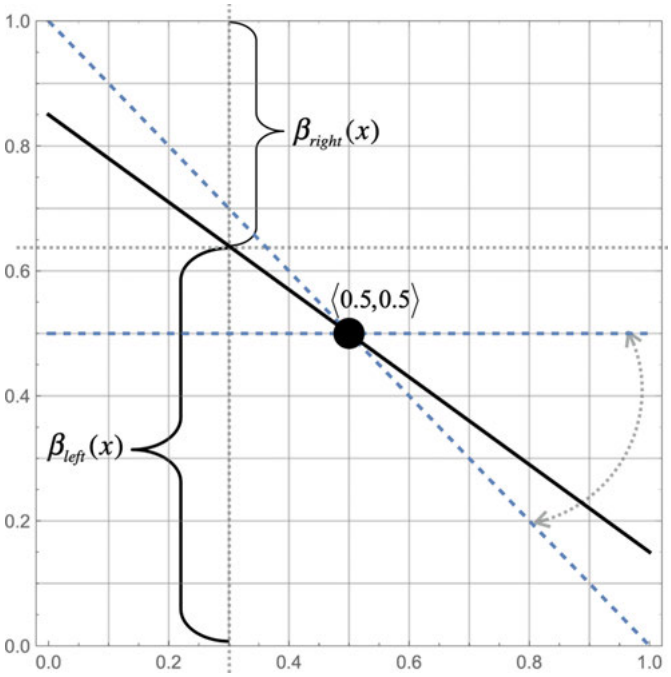


Figure 9.4: Opinion dependent bias.
Notes: The *x*-axis represents the opinions. The slope of the black graph gives the strength of the opinion dependent bias m (with $-1 \le m \le 0$) in equation (9.4). For an opinion x, $f(x)$ on the *y*-axis is the left proportion of the total size of the given confidence interval. $1-f(x)$ is the right proportion.

graph with $m = -0.7$, we get for an opinion $x = 0.3$ the proportions $\beta_l(0.3) = 0.64$ and, correspondingly, $\beta_r(0.3) = 0.36$. Note: whatever the value of m, for the center opinion $x = 0.5$ we always get $\beta_l(0.5) = \beta_r(0.5) = 0.5$.[11]

Figure 9.5, top, shows a polarization – two opinion camps, one to the left, one to the right of the center of the opinion space. It is a polarization based upon a *symmetric* confidence interval. As already noticed in Section 9.1.2, for certain regions of ϵ, polarization may evolve in the simple BC model. However, that is a rather *weak* polarization: the final distance between the two camps is only ≈ 0.36. The two camps are much closer to the center opinion 0.5 than they are to the extremes 0 and 1, respectively. A severe opinion dependent bias, for instance $m = -1$, produces polarization of another quality. Figure 9.5 shows the difference: The dynamics in Figure 9.5,

11 In Hegselmann and Krause (2002) we used the function $\beta_l(x) = mx + \frac{1-m}{2}$, defined $\beta_r(x)$ by $1-\beta_l(x)$, and used the range $m \in [0,1]$. As a consequence, the most severe bias is $+1$, while above it is -1. The only reason for the change is 'psychology': I find it more intuitive that a function over opinions 'from the left to the right' gives the *left* portion of the confidence interval. Anyhow, the old and the new approach are equivalent.

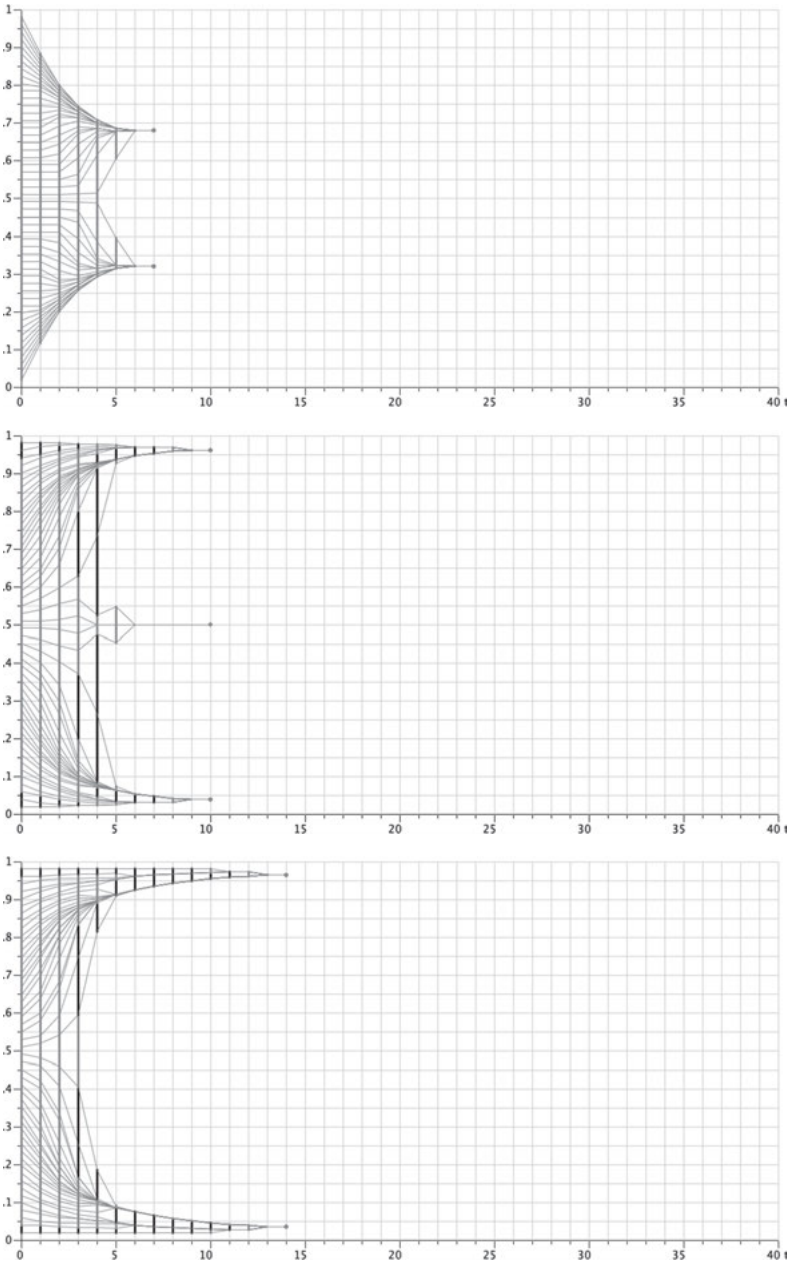


Figure 9.5: Effects of an opinion dependent bias.

Notes: *Top:* No opinion dependent bias, $m = 0$; $\epsilon_l = \epsilon_r = 0.2$ (symmetry). *Middle:* Opinion dependent bias $m = -1$, $\epsilon_{total} = 0.4$. *Bottom:* Bias $m = -1$, $\epsilon_{total} = 0.5$. Vertical *grey* lines indicate mutual influence between $x_i(t)$ and $x_{i+1}(t)$ in the profile (ascending order); black vertical lines indicate a *one-sided ϵ-split* in the opinion profile: only the more extreme of the two neighboring opinions influences the less extreme.

middle, and Figure 9.5, bottom, are both driven by the maximum bias $m = -1$. In both figures, the two opposite camps end up almost at the lower and upper bound of the opinion space: their distance is ≈ 0.92. Thus, a severe opinion dependent bias, produced a *strong* polarization: two camps at the opposite ends of the opinion space – and the center almost or entirely empty.

How comes? Two effects contribute. *First*, under symmetric confidence, those at the extremes are under a one–sided influence of less extreme agents. Their influence drives more extreme opinions direction center. With an increasing strength of the bias, the drive direction center *disappears*, or, depending upon the total size of the confidence interval, is significantly *weakened* at the extremes. (Whatever that size, under the maximum bias $m = -1$, the most extreme opinions, do not move at all.) The *second* effect works in the opposite direction. Under symmetric confidence it always holds: If agent i influences j, then j influences i as well. In contrast, under asymmetric confidence influence may be *one*-sided. And if so, then it is a very specific one-sidedness: only the less extreme opinion is influenced by the more extreme opinion. Together the two effects generate a drift to the extremes.

In Figure 9.5 vertical lines indicate the type of influence between neighbouring opinions.[12] Grey vertical lines between neighboring opinions indicate mutual influence between $x_i(t)$ and $x_{i+1}(t)$. Black vertical lines indicate one-sided influence; *either* only $x_i(t)$ influences $x_{i+1}(t)$ *or* only $x_{i+1}(t)$ influences $x_i(t)$. Wherever a vertical line is black, we have a *one-sided ϵ-split* in the opinion profile: only the more extreme of the two neighbouring opinions influences the less extreme – no moderating influence in the other direction. Thus, above 0.5, black vertical lines connect opinions for which only $x_{i+1}(t)$ influences $x_i(t)$; below 0.5, only $x_{i-1}(t)$ influences $x_i(t)$. Figure 9.5 demonstrates a remarkable point: Different from the two-sided ϵ-splits that are definitive splits for ever, one-sided splits may after some periods close again. In the meantime they generate an unmoderated pull to the more extreme position.

9.2.2 Strange effects

At this point, a generalising intuition suggests itself: As an opinion dependent bias m stepwise increases, a weak polarization under symmetric confidence will become more and more severe – monotonically increasing the distance between the two camps, that, at opposite ends, get closer and closer to the bounds of the opinion

12 Recall, the profiles are always ordered in an ascending order.

space. And indeed, for m-values from certain *partial sections* of $[-1, 0]$ that is true. But in general, the generalisation is definitively wrong.[13]

The refutation of the generalising intuition is given in Figure 9.6. As in the ϵ-diagrams in Figures 9.2 and 9.3, the y-axis shows stabilized end positions of opinions. Different from Figures 9.2 and 9.3, the y-axis represents the bias m. In steps of −0.01, the bias gets stronger and stronger. It starts with $m = 0$ (no bias at all) and ends with $m = -0.5$. Everything else (expected value start distribution with 50 agents, a confidence interval of a total size 0.4) is kept constant.[14]

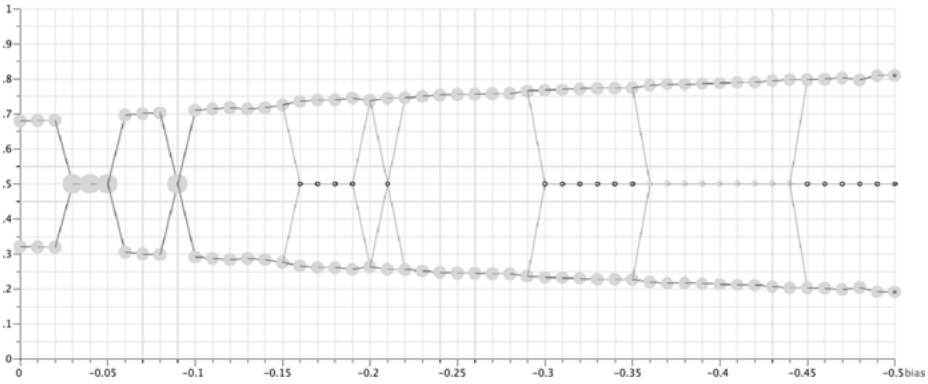


Figure 9.6: Stepwise stronger bias for the same start distribution.
Notes: On the x-axis the bias m gets stronger in steps of size -0.01. For a given bias, the y-axis indicates the stabilized end positions of opinions. The lines connect ranks in the ordered profiles. Different sizes of grey circles indicate the size of an opinion cluster. A small black circle with a white inner circle is a cluster of two opinions. All runs start with the same expected value start distribution $(n = 50)$. The total size of the confidence interval is always 0.4.

The most perplexing result is that an opinion dependent bias may lead to *consensus* when, for the same start distribution, a symmetric confidence interval of the same total size leads to polarization. For an expected value start distribution with $n = 50$ that is the case for $m = -0.03, -0.04, -0.05, -0.09$. Since all the underlying runs of Figure 9.6 are uniquely initialized, we can directly look into the single runs that produce the perplexing effect. Figure 9.7 does that for $m = -0.03$.

13 The relevant Section 4.2.2 of Hegselmann and Krause (2002) did (luckily enough) *not* claim that the generalisation is true, but – based upon that article and the underlying research – I believed the generalisation to be true until very recently.
14 We might refer to this type of diagrams as *ceteris-paribus*-diagrams (CP-diagrams). They show the final stable results of runs that are uniquely initialized; then everything is kept constant – except for *one* parameter.

The cause of the unexpected effect is visible in Figure 9.7: it is the evolution of a small cluster of just two opinions in the center of the opinion space. That cluster,

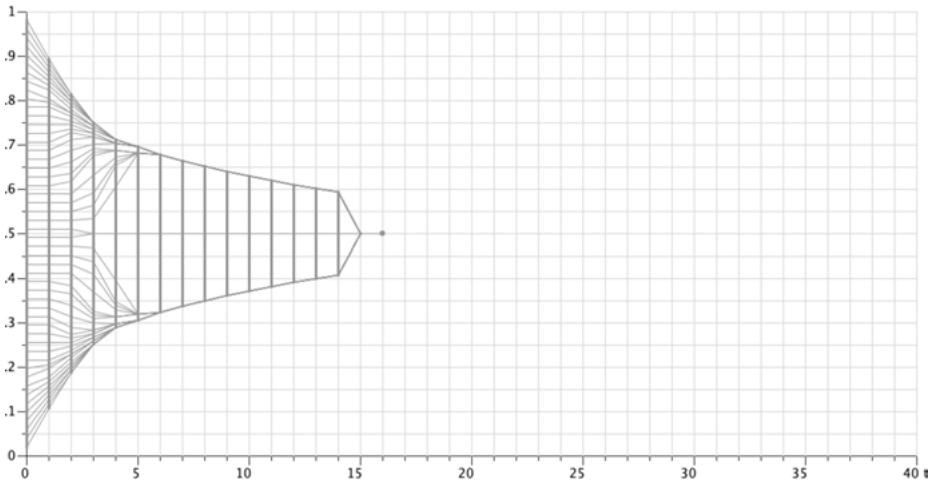


Figure 9.7: Consensus by biased confidence.
Notes: For an opinion dependent bias $m = -0.03$ the dynamics leads to consensus, while (for the same start profile and the same total size of ϵ) symmetry of the confidence interval leads to polarization (see Figure 9.5, top).

then, is a *bridge* between the two outer opinion camps and pulls them direction center. Nowhere in Figure 9.7 we see, indicated by a *black* vertical line, a one-sided ϵ-split – the perplexing results occurs without such splits. The same is true for all other cases of consensus in Figure 9.6.[15]

Careful inspection of Figure 9.6 reveals a further type of non-monotonicity: It is only a tiny margin, but the range of the final profile for $m = -0.20$ is *smaller* than for $m = -0.19$; in a certain sense, locally, a stronger opinion dependent bias has lead to a less extreme polarization. The same effect occurs for several other values of the bias m. In Figure 9.6 the bias gets stronger by a step size of –0.01. For both types of non-monotonicity (as to the number of finally stable clusters, and the final width of the profile), a smaller step size would unveil even more complicated non-monotonic structures. Obviously, an understanding of the polarization dynamics includes the task to understand surprising sensitivities and non-monotonicities.

15 In the range $m = -0.5, \ldots, -1$, I never found a case of consensus.

9.3 Extension two: A group of radicals enters the field

Let's now assume that there is a group of radicals. They all have the same opinion R, more or less close to the upper bound of the unit interval, for example $R = 0.9$, or even holding the most extreme position $R = 1.0$. (Alternatively, we might locate the radical position R close to the lower bound of our opinion space. Whatever we do, for the following it does not matter.) The radicals stick to their opinion – and that for ever[16]:

$$\mathrm{R(t+1) = R(t) = R.} \tag{9.5}$$

The size of the radical group may matter. We refer to the number of radicals by $\#_R$. Additionally we have the *normals*, i.e. agents as we know them from the simple BC model. They have opinions from the interval $[0, 1]$, they all have a strictly positive, constant, and symmetric $\epsilon > 0$. Normals update according to equation (9.2). Now the modification comes: Whenever the radicals are in a normal agent i's confidence interval, that is whenever $|x_i(t) - R| \leq \epsilon$, then *the whole group* of radicals is in $I(i, X(t))$. Since the radical group has $\#_R$ members, the radical position R is $\#_R$-times in $I(i, X(t))$.

Figure 9.8 shows a single run for 5 radicals with the radical position $R = 0.9$; the black horizontal line is their trajectory. They interact with 50 normals (expected value start distribution). The normals' confidence interval is $\epsilon = 0.2$. Without radicals,

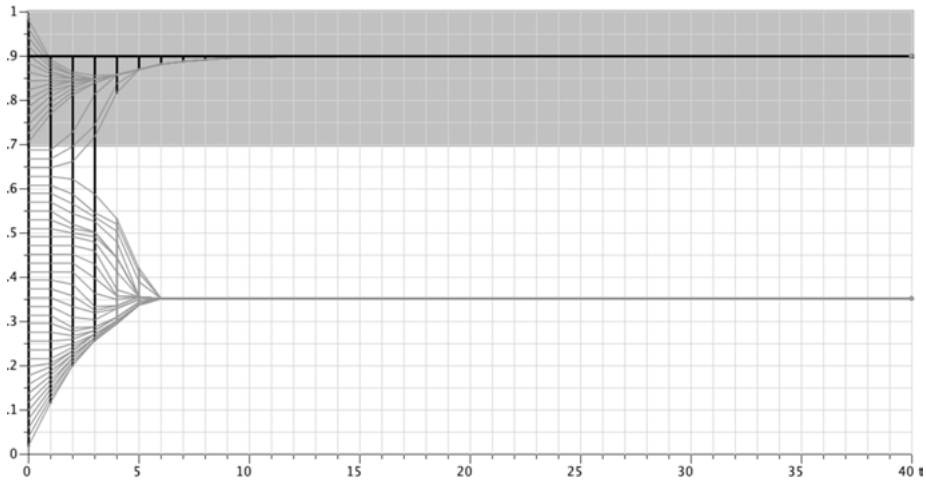


Figure 9.8: Normals and radicals.
Notes: 50 normals (expected value start distribution, $\epsilon = 0.2$) and 5 radicals ($R = 0.9$). Vertical black lines indicate a chain of direct or indirect influence of radicals on normals.

16 That is equivalent to assuming that their confidence interval is zero.

the opinion dynamics among the normals would be the dynamics in Figure 9.5, top. The dark grey area in Figure 9.8 indicates that part of the opinion space, in which all normals, given the size of their confidence interval, are under the *direct* influence of the radicals. Black vertical lines indicate the existence and length of a chain of direct *or indirect* influence of radicals on normals: Normals in the dark grey area are directly influenced by the radicals. But the radicals' influence does not end there. A normal j outside that area is indirectly influenced by a normal i inside the area of direct radical influence if $|x_i(t) - x_j(t)| \leq \epsilon$. Agent j, then, may influence other agents k outside the area of direct radical influence with opinions not further away than ϵ, and so forth. Figure 9.8 shows a far-reaching indirect influence of the radical group in the first periods: The chain of radical influence pervades the whole opinion profile, i.e. the radicals influence *all* normals. In period 4 that chain breaks. An upper part of the opinion profile converges towards the radical position. Below, the normals end up (obviously in finite time!) in a cluster. That cluster is far away from the radical position R. However, compared to the dynamics without the 5 radicals (see Figure 9.5, top), the lower cluster's final position is a bit shifted in the direction of the radical position. Obviously indirect radical influence matters.

It is very natural to think, that final numbers of radicalized normals crucially depend upon the number of radicals compared to the number of normals (hopefully only the ratio matters), the confidence level ϵ, and the radicals' position R. Under this working hypothesis, our model has only few parameters and we should be able to answer our questions for major parts of the parameter space. One possible simulation strategy, then, is the following: Let's assume we have 50 normal agents and the most extreme radical position that is possible, i.e. $R = 1.0$. Two parameters are left: The number of radicals, and the confidence level ϵ. Now we put a grid on the two dimensional parameter space: In 50 steps of size 0.01 the confidence level of normals increases from 0.01 to 0.5 on the x-axis. On the y-axis the number of radicals increases in 50 steps from 1 to 50 (then the group of radicals has as many members as the group of normals). We compute the runs for each of the 50×50 parameter constellations $\langle \epsilon, \#_R \rangle$. A run is considered stabilized, iff the opinion profiles $X(t)$ and $X(t+1)$ are almost the same. More precisely, we stop a run if for *all* agents i it holds that $|x_i(t+1) - x_i(t)| \leq 10^{-5}$. As to statistics, we will focus on *one* number only: the number of normals that finally hold an almost radical position. And we consider a normal agent i's opinion as almost radical iff $|x_i(t) - R| \leq 10^{-3}$.

Again, for all runs in the whole parameter space we will use one and the same expected value start distribution as given by equation (9.3). As a consequence, there isn't any randomness in our analysis. If we find in the grid of 50×50 parameter constellations $\langle \epsilon, \#_R \rangle$ some interesting effects, then we can go straight into the *unique single runs* that generated perplexing macro-effects.

There are other models of radicalization, for instance Deffuant et al. (2002) or Baurmann, Betz, and Cramm (2014). Often, they are more complicated than ours, or do not allow for the type of analysis that we want to apply. We go for extreme

simplicity and aim at a complete understanding, including the micro level that brings about the patterns on the macro level, for instance the numbers of finally radicalized agents.[17]

9.3.1 A bunch of surprising effects

Result of the computation of single runs for the 50×50 parameter constellations $\langle \epsilon, \#_R \rangle$ is an array of numbers. It is much easier to detect patterns and structures in *colored landscapes* rather than in an array of numbers (even if they are only integers – as in our case). Therefore (as a kind of phase diagram) Figure 9.9 shows, indicated by color, the number of normals that finally end up at the radical position $R = 1.0$. For ease of reference, we refer by the capital letters A, B, C, D, E, and F to certain regions of the parameter space, as they are partitioned by the black lines (two vertical, one horizontal).

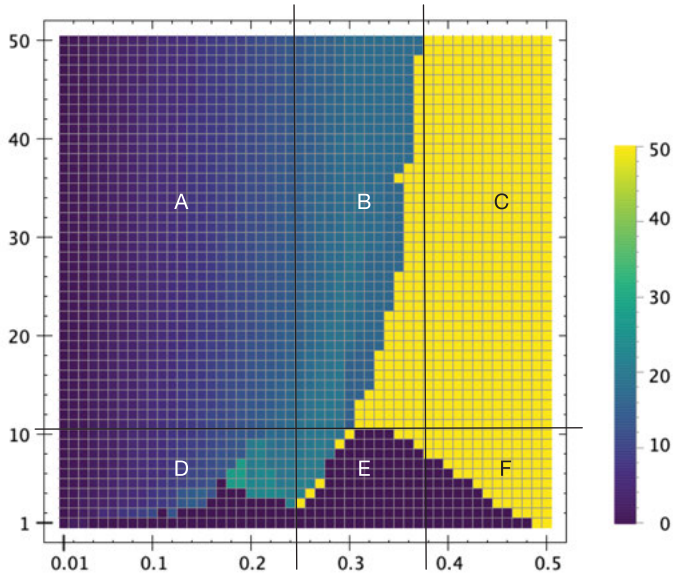


Figure 9.9: Radicalization landscape for an expected value start distribution.
Notes: *x*-axis: the confidence level increases in 50 steps of size 0.01 from 0.01 to 0.5. *y*-axis: the number of radicals increases from 1 to 50. Colors indicate the number of normals that end up at the radical position which is here assumed to be $R = 1.0$. The total number of normals is always 50.

17 The analysis presented here is extended in Douven and Hegselmann (2020).

On the *y*-axes the number of radicals increases stepwise. Therefore, sudden dramatic color changes in vertical direction are dramatic changes in the number of radicalized normals that – *ceteris paribus* – are caused by just *one* more radical. Correspondingly, a dramatic color change in horizontal direction is – *ceteris paribus* – a dramatic change in the number of radicalized normals caused by a tiny increase of ϵ by 1/100.

In the following we inspect region-wise our parameter space. We start with the three vertical regions. Our question is, how the number of normals that end up at the radical position, depends upon the number of radicals. Then an inspection of the two horizontal region follows. Our question there, is, how the radicalization of normals depends upon confidence levels.

1. In the region $F \cup C$, i.e. a region with *higher* confidence levels ϵ, the number of radicalized normals *monotonically* increases as $\#_R$ increases. But for all $\epsilon < 0.49$, there is a sudden jump: One more radical, and the number of radicalized normals jumps from none to all. Obviously there is an ϵ depending threshold $\#_R^*$ of radicals, such that, first, for that threshold *no* normal ends up at the radical position, while, second, for $\#_R^* + 1$ *all* normals end up at the radical position.

2. In the region $E \cup B$, i.e. a region with *middle-sized* confidence levels we find jumps of all sorts and in various directions: In region E there are – again in vertical direction – jumps from none to all: One more radical, and, instead of none, all normals end up at the radical position R. But, additionally, in region E and B there are jumps in the opposite direction: One *more* radical, and instead of all, *significantly less* (about half of the normals, or even less) end up radical.

By careful inspection of region $E \cup B$ in Figure 9.9 one can verify: With the exception of one of the ϵ values (the exception will be discussed later), it holds for the *middle-sized* confidence levels in region $E \cup B$:

1. For all of them exists an threshold $\#_R^*$ for none-to-all jumps.
2. For all of them exists another ϵ depending threshold $\#_R^{**}$ of radicals, such that, first, for that threshold *all* normals end up at the radical position, while, second, for $\#_R^{**} + 1$ *significantly less* normals become radical. Obviously, there is a second type of jumps, now working into the opposite direction.
3. In region E the second threshold $\#_R^{**}$ equals $\#^* + 1$. As a consequence, two steps of adding just one more radical causes the dramatic change from none to all, and then back to about half of the normals being radicalized.
4. For the thresholds $\#_R^{**}$ that are in region B, the jump from all to significantly fewer comes later – but it comes: There is always a number of radicals such that just one more reduces the number of radicalized normals from all to about less than the half.

There is one exception from observation (a) to (b): For step 28 on the *y*-axis ($\epsilon = 0.28$) there is a threshold $\#_R^{**}$, but no threshold $\#_R^*$. To be frank: I do not know

the reason. May be it is simply as it is – in the non-linear BC-dynamics often minor differences matter. The missing threshold may be a hint, that the exact position pattern of the thresholds $\#_R^*$ is a more complicated issue than it looks under our 50×50 grid of $\langle \epsilon, \#_R \rangle$ parameter constellations. A grid that is finer with regard to ϵ could give an answer. And finally, bleak as it is: The missing threshold $\#_R^*$ may be the consequence of numerical problems.[18]

Leaving the one exception aside and summing up: $E \cup B$, a region of *middle-sized* confidence levels, is a region with sudden ups (from none to all) and downs (from all to significantly less) radicalized normals. Along certain lines in the parameter space the sensitivity to tiny changes is extreme. The predominant phenomenon is, that the number of radicalized normals is *not* monotonically increasing with an increasing number of radicals. Just one more radical may lead to much less radicalization.

Even in the smooth areas of region B the radicalization of normals is clearly *not* monotonically increasing with regard to $\#_R$. On the contrary: In the left part of area B the radicalization of normals is slightly *decreasing* as the number of radicals *increases*.

3. The region $D \cup A$ is the region of *smaller* confidence levels ϵ. Again, for an increasing $\#_R$, there are certain threshold values where jumps occur. But they are not jumps from none to all. Nevertheless, they are jumps from none to a significant proportion. In the right part of D the sudden increase is more drastic than in the left part. Again there is a striking effect: Above the jumps from none to a significant proportion, the number of radicalized normals clearly *decreases* as the number of radicals increases – *less* radicals would have *more* effect.

4. *Horizontally*, i.e. with regard to $\#_R$, we distinguish *two* regions. There is an upper region with a *major* number (or proportion) of radicals, the region $A \cup B \cup C$. It is a region with always more than 10 radicals, i.e. a radical group size of more than $1/5$ of the number of normals, or, respectively, more than $1/6$ of the whole population.[19] Given such a major $\#_R$, if ϵ increases, there always exists a threshold ϵ^* such that for $\epsilon^* + 0.01$ the number of radicalized normals jumps from about $1/3$ to all. The upward jumps are compatible with monotonicity. However, careful color inspection of the area to the left of the thresholds ϵ^* clearly shows (especially clear for the middle sized confidence levels in region B) that, with increasing confidence levels, $\#_R$ – slightly and smoothly – first increases, but then decreases (again slightly and smoothly). In sum, for major numbers of radicals, as to the numbers of radicalized normals, there is no general monotonicity with respect to the size of their confidence interval.

18 For a description and discussion of the numerical problems see Hegselmann and Krause (2015: 483ff.).

19 If one counts the cells up to the horizontal line, the result is 10. Note: The y-axis' origin is 1 (and the x-axis origin is 0.01).

There are *two* regions in the whole parameter space, that behave very smoothly: the regions A and C. Both belong to the upper horizontal region with a *major* number of radicals that we inspect right now. In region C, that is for *higher* confidence levels, for any $\#R$ all normals end up radicalized. In region A, that is for *smaller* confidence levels, it is the normals' confidence level that matters – not $\#_R$: The radicalization of normals increases as their ϵ increases. In the bottom right area of A the number of the radicals has a bit effect: The number of normals, that end up at R, slightly decreases as $\#_R$ increases.

However, a warning side remark: Figure 9.9 shows the number of normals that become radical. We consider a normal agent i as "radical", "radicalized", "ending up at the radical position" etc. iff $|x_i(t) - R| \leq 10^{-3}$. Therefore, even if $\#_R$ has (almost) no effect on the number of – in this sense – radicalized normals, it may nevertheless have (and often has) a major effect on the mean or median opinion of the normals' opinions, the cluster structure etc., which we do not analyze here.

In region A and C the number of radicals has very little or no effect on the radicalization of normals. In region B that is different, and $\#_R$ seriously matters: The exact location (though not the existence) of the threshold $\epsilon^\star$ depends upon the number of radicals: As their number increases, the jumps occur more to the right, i.e. they require higher confidence levels.

5. The lower horizontal region, i.e. $D \cup E \cup F$, is a region with *minor* numbers of radicals (not more than $1/6$ of the whole population of normals *plus* radicals). In terms of jumps it is the wildest region: In E and F we find (as in B) thresholds ϵ^* such that for $\epsilon^* + 0.01$ the number of radicalized normals jumps from less than a half or even none to all. But, additionally, there are values ϵ^{**}, such that for $\epsilon^{**} + 0.01$ the number of radicalized normals jumps from all to zero. The most striking point is, that in E both threshold values are horizontally next to each other, i.e. $\epsilon^{**} = \epsilon^* + 0.01$. Obviously, the radicalization of normals reacts in E, i.e. an area with both, a minor number of radicals *and* a middle-sized confidence level, extremely sensitive with regard to both initial conditions, the confidence level *and* the number of radicals.

However, there is again a conspicuity in area E, now in vertical direction: For all $\#_R > 2$ *except* for $\#_R = 6,\ 7,\ 8$ there exists a threshold ϵ^* (as defined above). Again we do not know the reason. The same considerations, as mentioned above in the corresponding case for $\epsilon = 0.28$, apply (see the second observation).

In area D we find for an increasing ϵ a lot of jumps in both directions: from some to none and from none to some. The jumps are less dramatic than in region E, but they are there. For a description one might introduce thresholds that correspond ϵ^* and ϵ^{**} but have reduced requirements.

To sum up, with respect to the confidence level, the whole region $D \cup E \cup F$ is a region of non-monotonicity, jumps up and jumps down.

What are the main results of our inspection? There are two 'smooth' areas in the parameter space, the areas A and C. But in *all* other areas we find the thresholds ϵ^*,

ϵ^{**}, $\#_R^*$, and $\#_R^{**}$ (in D we observe corresponding thresholds with reduced requirements). Let's call a region of our partitioned $\langle \epsilon, \#_R \rangle$-parameter space *wild* iff, first, we have in that region a non-monotonicity (both, decreasing *and* increasing) with respect to one or both parameters, and, second, the region is pervaded by sensitivities. Under that definition we can distinguish *two* wild regions: In *vertical* direction the region $E \cup B$, a region of middle-sized confidence levels; in *horizontal* direction the region $D \cup E \cup F$, a region of comparatively small numbers of radicals.

For all regions, wild or not, immediately "Why is it, that . . . ?"- questions arise. Why is it, that in region C neither the number of radicals, nor the confidence level has any effect on the number of radicalized normals? Whatever the specific parameter constellation in that region, *all* normals end up radical – but why? Why is it, that in region A, a region where the number of radicals is above $1/6$ of the whole population, radicalization of normals is *not very much* influenced by the number of radicals. Obviously, it is the confidence level of normals that matters – but why?

In the following we will make *expeditions* into the 'wild' regions of our parameter space and try to understand, how 'in the deep' some completely deterministic mechanisms create the wild '*radicalization landscape*', that Figure 9.9 displays.

9.3.2 An expedition into the wild parameter region

Our use of always the same expected value start distribution has *two* consequences: First, whenever we want to understand what causes certain effects in our radicalization landscape, we can go directly into *unique single runs*. No statistical analysis of 100 or so randomly started runs is necessary. Second, since, additionally, all the single runs start with the same start distribution of normals, we can, by comparison of single runs, *on the level of single agents* directly observe the effects of changes of ϵ or $\#_R$. Especially, if we inspect single-run-sequences of *small stepwise changes*, we directly observe the working of the 'forces in the deep' that generate the surface of our radicalization landscape – and that should be a good starting point for an identification and understanding of the mechanisms that bring about the puzzling landscape.

Our expedition will be an expedition direction north that starts at $\epsilon = 0.31$. On that path are two puzzling points: First, a sudden change from none to all of the normals being radicalized (explanandum 1); second, with only a few more radicals, a sudden change from all to only about $1/3$ normals being radicalized (explanandum 2). For each of the parameter constellations that we pass going north on the 50×50 grid of $\langle \epsilon, \#_R \rangle$, we have unique single runs with the same start distribution. Therefore, we can generate a sequence of 50 pictures, one for each $\langle \epsilon, \#_R \rangle$ constellation that we pass. Each of the pictures displays the trajectories of all 50 agents. The time scale on the x-axis is always the same: 50 periods. Going strictly north implies, that ϵ is kept constant. Therefore, whatever is changing in the sequence of pictures, it is the consequence of *one* factor only: the number of radicals.

Figure 9.10 displays the *first* explanandum: The jump from none to all normals being radicalized when the number of radicals increases from 10 to 11. For an explanation we start in Figure 9.10 bottom, i.e. the dynamics under the influence of 10 radicals. As the dark grey vertical lines indicate, there is up to period 7 (we start with period 0) a chain of direct or indirect influence of the 10 radicals even on the most distant normals. Soon 3 opinion clusters emerge among the normals. The cluster in the middle functions as a bridge between the upper and the lower cluster. The upper cluster is in the dark grey area of the opinion space, and that is the area of direct influence of the radicals. Thus the upper cluster of normals is a bridge between the radical group and other two clusters of normals, which, period by period, move direction *R*. But that works only for a while: As the lower cluster of normals moves

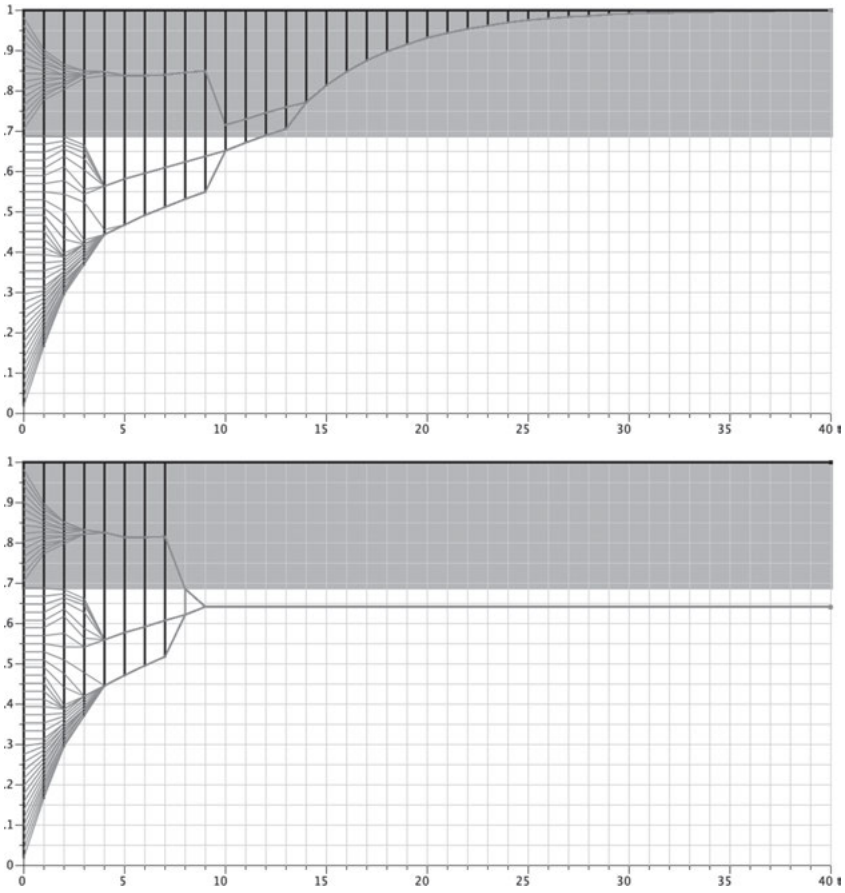


Figure 9.10: One more radical and everybody gets radicalized.
Notes: *Bottom:* $\#_R = 10$. *Top*: $\#_R = 11$. $\epsilon = 0.31$. One more radical causes a jump from none to all of the normals ending up at the radical position.

upward, at a certain point the upper normal cluster has both, the middle and the lower cluster within its confidence interval. Their combined influence on the upper cluster is strong enough to pull the upper cluster completely out of the area of direct influence of the radicals. The consequence is, that no bridge between radicals and normals exists any longer. However, the radicals had an effect on the normals: Without radicals the normals would end up at a 0.5-consensus. With 10 radicals (1/5 of the number of normals, 1/6 of the whole population) it is about 0.64.

Now we add just one radical, analyze the trajectories in Figure 9.10 top, and compare it with what we see in Figure 9.10 bottom: With the one more radical, again, the three clusters of normals evolve. However, it takes a few periods more until the upper cluster gets under the direct influence of both clusters below, which, therefore, both are moving direction *R* a bit longer, before, then, that period comes. When it comes, the upper cluster makes – as in the case with one radical less – a steep move away from the radicals' position. But different from the case with $\#_R = 10$, the upper cluster does *not* get out of the area of direct influence of radicals. From that moment onwards everything is lost: Though now further away from the radical position, the upper cluster *continues* to function as a bridge between all other normals and the radicals. After two more periods the bridge isn't necessary any longer: Now *all* normals are under the *direct* influence of the radicals – and that is a point of no return: From now on (and as one cluster) all normals irreversibly move and converge – though in infinite time – to the radical position.

The inspection of the underlying dynamics resolves the puzzling effect in the radicalization landscape: The sudden jump from none to all is due to what we might call a *positive bridging effect*: For $\#_R = 10$ a cluster evolves that for a while functions as a bridge between the radicals and all other normals. But then the bridge breaks down and, additionally, the upper cluster gets out of the area of direct radical influence. For $\#_R = 11$ the bridge to the radicals *continues* to function until it becomes superfluous. The functioning of the bridge is critical and accounts for the difference between none or all of the normals ending up radical.

Figure 9.11 displays the *second* explanandum: The sudden *drop down* from all normals being radicalized to only 1/3, and that by *increasing* the number of radicals from 13 to 14. We start our analysis in Figure 9.11 bottom. What we see there, is very similar to Figure 9.10 top. Based on what we saw there, we understand the positive bridging effects that are (still) at work in the dynamics in Figure 9.11, bottom. But the one more radical in Figure 9.11, top, causes dramatically different trajectories: the middle cluster of normals is somehow 'blown up': One of the former members joins the upper cluster, all others join the lower cluster. For some periods the radicals still influence even the most distant normals. But there is no evolution of a bridging cluster *in-between* the two clusters of normals. The distance between the two clusters enlarges. As a consequence, the radicals' chain of influence breaks and becomes very short afterwards. More than 2/3 of the normals form a cluster at about 0.45, that is even slightly *lower* than the center of the opinion space.

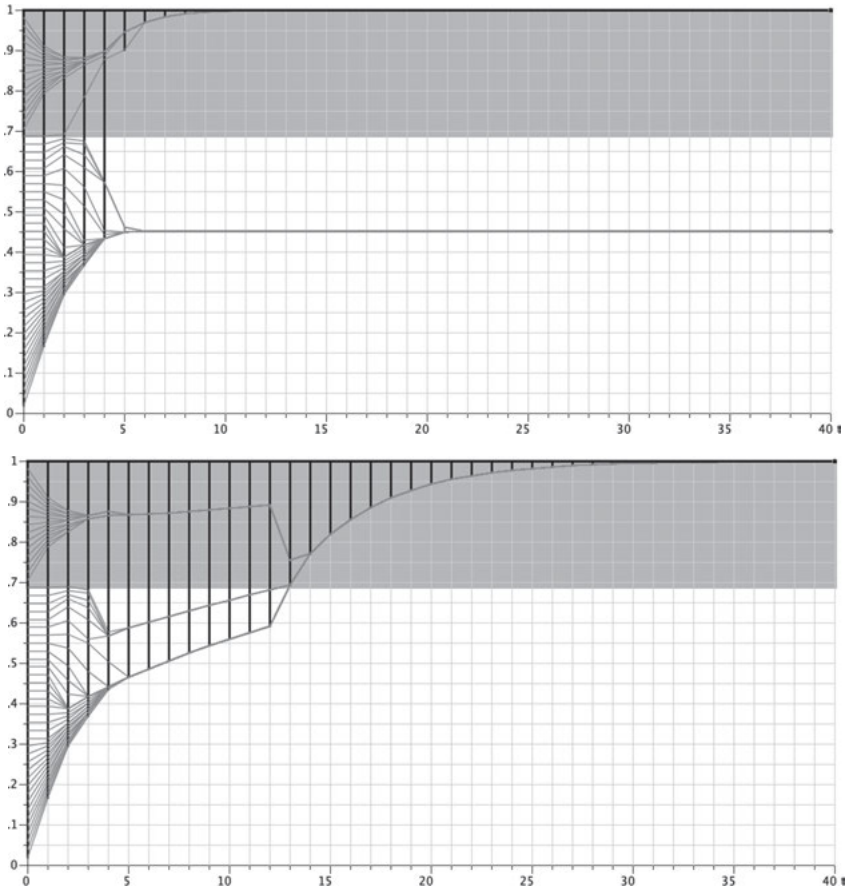


Figure 9.11: One more radical and the radicalization goes down.
Notes: *Bottom*: $\#_R = 13$. *Top*: $\#_R = 14$. $\epsilon = 0.31$. One more radical causes a jump down from all to less than one third of the normals ending up at the radical position.

Decisive for the sudden jump downwards is, that the one more radical causes a rupture in a segment of the normals' opinion profile that, without the additional radical, would have become a bridging cluster. – Obviously, there are not only positive bridging effects. In our explanation of the second explanandum a *negative* bridging effect is at work: Under one more radical a former bridge to the radicals ceases to exist. And that causes a dramatic reduction in terms of radicalized normals.

We can't analyze here all the details of the mechanics that works underneath the radicalization landscape on the route further north. Only so much: The number of radicalized normals goes up to a maximum of 19, fluctuates for a while between 18 and 19, and ends for $\#_R = 50$ with 18 radicalized normals. *All* explanations of these figures, their small range, and their fluctuations, are about the details of how the cluster, that for $\#_R = 13$ functions as a bridge between the upper and the lower

cluster of normals (see Figure 9.11, bottom), is 'blown up', disassembled, and ruptured into pieces, once we add one more radical, and another one, and so forth.

In a similar style we could do an expedition going east, for instance starting at $\#_R = 5$. As Figure 9.9 shows, on that path we would encounter three dramatic sudden changes that ask for explanation.[20] What lessons can we learn from our expeditions? The most important lessons are these:

1. A prominent role in all explanations of sudden jumps of the number of finally radicalized normals have *bridges from* normals *to radicals*. They require, as a kind of pier, normals (cluster or single) that, given a confidence level ϵ,
   1. are themselves *inside* the area of direct radical influence,
   2. are *within* the confidence interval of *other* normals, that are *outside* the area of direct radical influence.

Let's call such bridges *type-R bridges*. They are decisive for any influence of radicals outside their limited area of direct influence (which is determined by the normals' confidence level). Type-*R* bridges allow for *indirect* influence of radicals on normals.

There is a second type of bridges, *bridges from* normals *to* normals. They require (again as a kind of pier) normals (cluster or single), that, given a confidence level ϵ,

1. are themselves *inside* the confidence interval of at least *two other* normals (cluster or single),
2. that themselves are *outside* each other's confidence interval.

Let's call this type of bridge *type-N bridge*.

What we have seen, then, is, that via an uninterrupted chain of bridges, starting with a type-*R* bridge and then prolongated by a number of type-*N* bridges, the radicals may have an influence even on normals that are far away from the radical position. But, except for the group of radicals, the piers of our bridges *can move over time* – and that may *destroy* a bridge, whether of type-*R* or type-*N*. At the same time *new* piers for *new* bridges may evolve. *Lesson:* Understanding the radicalization landscape is an understanding of *types, evolution, and breakdown of bridges in a dynamical network*.

2. The probably most striking puzzles are sudden jumps *down* from all to none, or significantly fewer radicalized normals – and that caused by an *increasing* $\#_R$ or ϵ (as an expedition direction east would demonstrate). In explanandum 2 one more radical causes a pull upwards, which, *via* a type-*R* bridge, disrupts a former and essential type-*N* bridge. The movable pier of the type-*N* bridge moves steeply direction *R*. Thereby the bridging capacity, given by ϵ, is over-stretched, and the type-*R*-bridge breaks down (see Figure 9.11).

20 Hegselmann and Krause (2015) describes that expedition in detail.

An increasing number of radicals may have the effect that the upward pull disrupts piers of type-*N* bridges and/or attracts too fast the pier of a type-*R* bridge. An increase of ϵ causes a stronger contraction. That may sweep along a former pier of a type-*R* bridge, and the former pier gets outside the area of direct radical influence. In both cases the breakdowns of bridges depends upon thresholds. Therefore they are sudden events. In both cases the breakdown of bridges may stop the radicals' influence on major fractions of normals. As to the numbers of radicalized normals, even jumps from all to none are possible. *Lesson:* We can explain the sudden jumps downward in terms of effects on type-*R* and type-*N* bridges in an opinion profile, that is exposed *to two interlinked forces*, that get stronger: the first pulls upwards, the second contracts the range of the profile. If the forces get *stronger*, they may *destroy* decisive bridges of influence.

3. Sudden jumps upwards are another puzzling effect. Such a jump occurs in the *explanandum* 1. With one more radical the pier of a type-*R* bridge is no longer pulled downwards outside the area of direct radical influence (see Figure 9.10). *Lesson:* Obviously we can explain the sudden jumps downwards in terms of effects of the interlinked forces for type-*R* or type-*N* bridges. If the forces get stronger, bridges that before broke down, may keep functioning, or piers for new bridges may evolve.

That are some lessons. Many questions are left open, for instance with regard to the location of the extreme sensitivities. To answer them requires many more expeditions into the wild.

9.4 Concluding remarks and future perspectives

In this article we presented two extensions of the BC Model. In Section 9.2 we introduced an opinion dependent bias. In Section 9.3 radicals entered the field. Both extensions revealed phenomena of general importance: *First*, the importance of bridges; *second*, an extreme sensitivity in certain regions of the parameter space. In these regions very small differences in the parameter values have massive effects.

With regard to *bridges*, in the polarization processes of Section 9.2, opinion clusters that had developed in the middle of the opinion space were able to pull evolving outer clusters towards the center; without the bridge in the middle, the outer clusters would belong to different opinion worlds. Conversely, the absence or disappearance of such bridges contributed decisively to polarization. In Section 9.3, bridges or chains of bridges were decisive for the influence of a group of radicals on normal BC agents. They connected the cluster of radicals with clusters of normals outside the direct influence of the radicals. The ϵ splits that we introduced in the context of the basic BC model, can also be described as the absence or disappearance of bridges. The term *bridge* in the sense as we use it here, originates from network theory. The term is explicitly introduced in the network theory classic Harary,

Norman, and Cartwright (1965). In the classical sociological application of network theory, namely in Granovetter's *The Strength of Weak Ties* (1973), the weak ties *are* bridges. The use of the term in the context of the BC model is *not* a transfer of the concept into another field: The BC dynamics can be understood as a dynamic network in which the edges connect agents (nodes) whose opinion distance is not greater than ε.

As to *sensitivity*, we saw in the first extension that polarization reacts extremely sensitive to small changes of a relatively weak bias. In the second extension the number of radicalized normals reacts extremely sensitive to small changes of a relatively small proportion of radicals or, respectively, small changes of a medium-sized confidence interval. The sensitivity is associated with certain non-monotonicities: a more pronounced bias does not necessarily lead to a more pronounced polarization – it can even lead to consensus; while an even stronger bias then leads to polarization again. In the second extension, more radicals or larger confidence intervals can lead to a larger, but also to a smaller number of radicalized normals.

In the first instance, this sensitivity is the sensitivity of artificial worlds created as instantiations of the BC model. If, however, with regard to relevant and critical aspects, these artificial worlds are sufficiently similar to the real world, then one would also have to reckon with all the mechanisms and effects in the real world. As a consequence, for an understanding of polarization and radicalization processes in the real world, a concentration on network bridges would be a good idea. As far as details are concerned, things could become epistemically and politically difficult in the real world: In certain parameter ranges, anyone who wants to make a prognosis on the results of an actual dynamics of opinions would have to have extremely precise knowledge of the situation, that is the actually given parameter values. The same applies to the exact explanation of a factual dynamics. A social planner who could within certain but narrow limits, manipulate parameter values, would have to know very precisely where he or she actually is in the parameter space in order to even give his or her intervention only the desired direction. In a radicalization landscape that we calculate for one and the same *random* start distribution, the difficulties would become even more severe.

All results and insights of this article are based upon one and the same expected value start distribution of 50 opinions according to equation (3). That made it easy to analyze the factors and mechanisms that cause sudden massive changes in sensitive parameter regions. But are the findings that the radicalization landscape of Figure 9.9 shows perhaps essentially due to the *equidistance* of opinions in the expected value start profile that deterministically idealizes an even random start distribution? Would this landscape have looked substantially different if one carries out 50, 100 or 1000 experiments with 50 randomly distributed opinions of normals (uniform distribution) and then averages over the number of radicalized normals? The answer is given by such random simulations carried out by Igor

Douven.[21] As Figure 9.12 shows, the answer is: *No*, the radicalization landscape calculated in this way does not look much different. The abrupt changes are slightly blurred, but they remain abrupt – the sensitivity is preserved. In this respect, the general use of the same expected value start distribution is obviously a good representative substitute for high numbers of random experiments. The required computing time is reduced by one to three orders of magnitude. A radicalization landscape as the 50×50 parameter constellations $\langle \epsilon, \#_R \rangle$ with colors indicating the final number of radicalized normals, constitutes a major *explanandum*. If the underlying start distribution is both representative and always the same, there is a straightforward strategy to find explanations: we directly go into the single runs that produced the puzzling radicalization patterns. In our case the single run analysis made very clear that the crucial point is the emergence and the collapse of bridges (type-*R* or type-*N* bridges) between clusters.

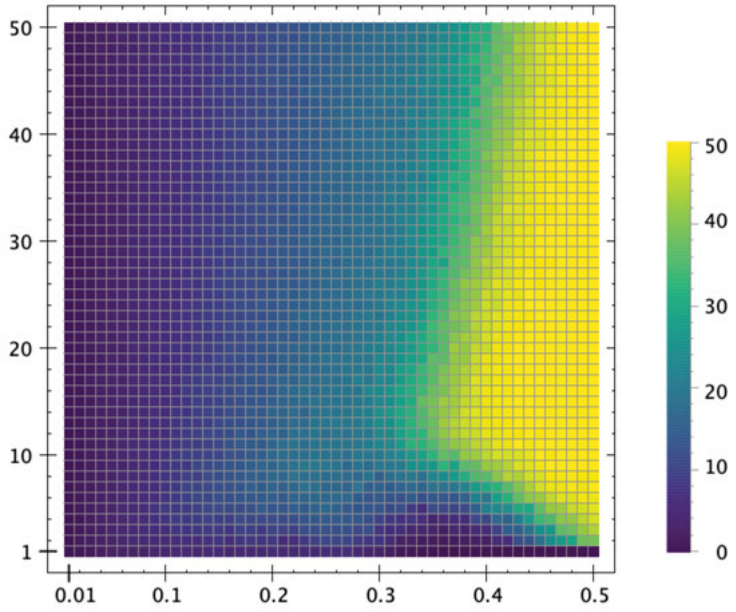


Figure 9.12: Radicalization landscape for random start distributions.
Notes: For each of the 50×50 parameter constellations $\langle \epsilon, \#_R \rangle$, Igor Douven ran 100 simulations based upon random start distributions (uniform). *x*-axis: the confidence level increases in 50 steps of size 0.01 from 0.01 to 0.5. *y*-axis: the number of radicals increases from 1 to 50. As in Figure 9.9, the radical position is $R = 1$. Colors indicate the *average* number of radicalized normal.

21 On November 1, 2018, Igor Douven has sent to me the respective radicalization landscapes. They were computed with the newly developed, very fast *Julia* programming language.

All that does not mean that, from now on, we should forget about iterated runs based upon random start distributions. In the research strategy that I propose here, runs with random start distributions are still important, but more as a *hedging measure*: By computing as well the radicalization landscape for iterated random start distributions we hedge against the danger of artifacts, as they may be caused by expected value start distributions. The similarity between the landscape that is based upon one and the same expected-value start distribution (Figure 9.9) and the landscape that is based upon iterated runs based upon random start distributions (Figure 9.12) suggests that our single-run based explanations of puzzling phenomena regard and cover generic effects.

References

Baurmann, Michael, Gregor Betz, and Rainer Cramm. 2014. "Meinungsdynamiken in fundamentalistischen Gruppen – Erklärungshypothesen auf der Basis von Simulationsmodellen." *Analyse und Kritik* 36 (1): 61–102.

Deffuant, Guillaume, David Neau, Frederic Amblard, and Gérard Weisbuch. 2000. "Mixing beliefs among interacting agents." *Advances in Complex Systems* 3 (01n04): 87–98.

Deffuant, Guillaume, Fréderik Amblard, Gerald Weisbuch, and Thierry Faure. 2002. "How can extremism prevail? A study based on the relative agreement interaction model." *Journal of Artificial Societies and Social Simulation* 5 (4).

Douven, Igor. 2010. "Simulating peer disagreements." *Studies in History and Philosophy of Science* 41 (2): 148–157.

Douven, Igor and Kelp, Christoph. 2011. "Truth Approximation, Social Epistemology, and Opinion Dynamics." *Erkenntnis* 75: 271–283.

Douven, Igor and Sylvia Wenmackers. 2017. "Inference to the Best Explanation versus Bayes's Rule in a Social Setting." *British Journal for the Philosophy of Science* 68: 535–570.

Douven, Igor and Rainer Hegselmann. 2020. "Mis- and Disinformation in the Bounded Confidence Model." Submitted.

Flache, Andreas, Michael Mäs, Thomas Feliciani, Edmund Chattoe-Brown, Guillaume Deffuant, Sylvie Huet, and Jan Lorenz. 2017. "Models of Social Influence: Towards the Next Frontiers." *Journal of Artificial Societies and Social Simulation* 20 (4).

Granovetter, Mark. 1973. "The Strength of Weak Ties." *American Journal of Sociology* 78 (6): 1360–1380.

Harary, Frank, Robert Z. Norman, and Dorwin Cartwright. 1965. *Structural Models. An Introduction to the Theory of Directed Graphs*. New York: John Wiley & Sons, Inc.

Hegselmann, Rainer and Ulrich Krause. 2002. "Opinion dynamics and bounded confidence: models, analysis and simulation." *Journal of Artificial Societies and Social Simulation* 5 (3).

Hegselmann, Rainer and Ulrich Krause. 2015. "Opinion dynamics under the influence of radical groups, charismatic leaders, and other constant signals: A simple unifying model." *Networks & Heterogeneous Media* 10 (3): 477–509.

Hegselmann, Rainer, Stefan König, Sascha Kurz, Christoph Niemann, and Jörg Rambau. 2015. "Optimal Opinion Control: The Campaign Problem." *Journal of Artificial Societies and Social Simulation* 18 (3).

Krause, Ulrich. 1997. “Soziale Dynamiken mit vielen Akteuren. Eine Problemskizze.” Pp. 37–51 in *Modellierung und Simulation von Dynamiken mit vielen interagierenden Akteuren*. Edited by Ulrich Krause and Manfred Stöckler. Bremen: Universität Bremen.

Krause, Ulrich. 2000. “A discrete nonlinear and non-autonomous model of consensus formation.” Pp. 227–36 in *Communications in Difference Equations. Proceedings of the Fourth International Conference on Difference Equations, Poznan, Poland, August 27–31, 1998*. Edited by Saber Elaydi, Gerry Ladas, Jerzy Popenda, and Jerzy Rakowski. Amsterdam: Gordon and Breach Science Publishers.

Lehrer, Keith and Carl Wagner. 1981. *Rational consensus in science and society. A philosophical and mathematical study*. Dordrecht: D. Reidel Publishing Company.

Lorenz, Jan. 2007. “Continuous opinion dynamics under bounded confidence: a survey.” *International Journal of Modern Physics C* 18 (12): 1819–1838.

Pariser, Eli. 2011. *The Filter Bubble: How the New Personalized Web Is Changing What We Read and How We Think*. New York: Penguin Press.

Sîrbu, Alina, Vittorio Loreto, and Vito D. P. Servedio. 2017. “Opinion Dynamics: Models, Extensions and External Effects.” Pp. 363–401 in *Participatory Sensing, Opinions and Collective Awareness*. Edited by Vittorio Loreto, Muki Haklay, Andreas Hotho, Vito D.P. Servedio, Gerd Stumme, Jan Theunis, and Francesca Tria. Cham: Springer International Publishing.

Sunstein, Cass R. 2017. *#republic: Divided Democracy in the Age of Social Media*. Princeton, N.J.: Princeton University Press, 2017.

Urbig, Diemo, Jan Lorenz, and Heiko Herzberg. 2008. “Opinion Dynamics: the Effect of the Number of Peers Met at Once.” *Journal of Artificial Societies and Social Simulation*, 11 (2).

Wedin, Edvin and Hegarty, Peter. 2015. “The Hegselmann-Krause Dynamics for the Continuous-Agent Model and a Regular Opinion Function Do Not Always Lead to Consensus.” *IEEE Transactions on Automatic Control* 60 (9): 2416–421.

Xia, Haoxiang, Huili Wang, and Zhaoguo Xuan. 2011. “Opinion Dynamics: A Multidisciplinary Review and Perspective on Future Research.” *International Journal of Knowledge and Systems Science* 2 (4): 72–91.

Michał Bojanowski

10 Local Brokerage Positions and Access to Unique Information

Abstract: The theory of structural holes (Burt 1995) provides an explanation regarding network positions and their associated benefits in many social settings. One of the key points of the theory is formulated on the node level and is about efficient access to information and resources. Namely, actors bridging structural holes by connecting otherwise disconnected segments of a social network, the brokers, have access to information that circulates over the network while simultaneously maintaining relatively small number of ties. For example, scientists often maintain non-redundant collaboration ties because they are associated with academically-relevant resources unavailable in other collaborations. Apart from the above node-level property one can formulate a tie-level property corresponding to the 'tie redundancy'. Studying a process of the network information diffusion explicitely with a computational experiment and tools of information theory allows us to compare local properties of the diffusion with structural features related to the Burtian notions of brokerage and redundancy. We argue and demonstrate that (1) the above mentioned node-level and tie-level properties can be supplemented with a triad-level one. Specifically, that the notion of 'tie redundancy' should be in fact considered as an essentially triadic property. Further, (2) each of the three properties is associated with a viable empirical strategy of testing whether (aspects of) the structural holes mechanism is at work in a given network.

10.1 Introduction

The theory of structural holes (Burt 1995) identifies network positions having ties that connect otherwise unconnected segments of a social network as brokerage positions. Occupying a brokerage position in social networks is usually associated with various benefits. Further, the theory stipulates, among other things, that it is beneficial for an actor to occupy such broker positions as it allows for efficient access to information that spreads in the network. This statement pretends to the node level – it specifies that certain actors (network nodes), namely brokers, will be more successful than other actors given some appropriate notion and measure of success. For example, managers occupying such positions in an organization have been shown to be more likely to be promoted (Burt, Hogarth, and Michaud 2000).

Note: Author thanks Polish National Science Centre (NCN) for support for the RECON project (http://recon.icm.edu.pl/) through grant 2012/07/D/HS6/01971.

Michał Bojanowski, Kozminski University

https://doi.org/10.1515/9783110647495-010

Among the other elements of the theory of structural holes is the one about information flow and access to resources. It is related to the notion of *tie redundancy*. A tie is redundant for an actor to an extent that it connects others, to whom an actor already has access via his other ties. This concept is related to the concept of *weak ties* (Granovetter 1977). Granovetterian "weak ties", or Burtian "non-redundant" ties, are usually sources of information that is less likely to be acquired through "strong" or "redundant" ties. As such, it is a statement on the tie level. It has been famously shown by Granovetter (1974) that people usually learned about their current job through social contacts who are not well embedded in their personal networks.[1]

When approaching a research setting with a question whether the structural holes mechanism is at work one usually tests one of the two implications mentioned above: whether brokers are more succesful than non-brokers or whether relevant information is usually coming to an actor through a non-redundant rather than a redundant tie. Consider Figure 10.1 presenting a hypothetical personal network of actor *A*. We assume that this personal network is embedded in a much larger network – alters of *A* might have further connections, but we do not know them. We can (1) mark through which actor *A* learned the information of interest (e.g. about a job), and (2) measure the extent of redundancy of each of *A*'s ties, e.g. using *dyadic redundancy*. If the tie-level implication of structural holes is true the information is more likely to come to *A* from alters with low redundancy. Most likely actor *F*.

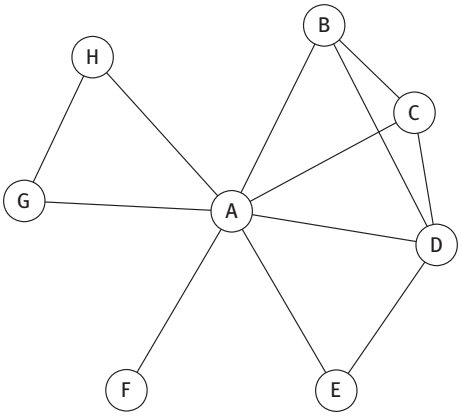


Figure 10.1: For actor A alters G, H, and E are redundant to the same extent. At the same time, sets of actors G-H and B-C-D-E can be sources of different information as they are disconnected from each other.

1 While the issues addressed in this paper can be argued to operate on a scale of complete, as opposed to ego-centric, graphs (see Borgatti 2005) we are limiting our attention to personal networks. This is of great relevance as a lot of empirical studies of the mentioned phenomena, including the seminal work of Granovetter (1974), is conducted with ego-centrically sampled network data.

We argue that the abovementioned approach is often insufficient. Consider a research setting in which there are multiple types of information circulating in the social network and none of these types is any way more special than other types. For example, collaboration between scientists is often motivated by a possibility to acquire or exchange various resources relevant for functioning in academia, such as specific expertise, access to costly equipment or mentorship.[2] As all types of information are equally salient we cannot approach the problem, as suggested by the tie-level implication, by studying a probability of receiving "new" information as a function of tie redundancy. None of the information is "new". What we need instead is to analyze whether the information received through a non-redundant tie is *different* from the information recieved through the other ties. In Figure 10.1 actors E, G, and H are redundant for the actor A to the same extent because each has exactly one tie to other contacts of A. However, they are clearly not redundant *vis a vis each other*. For example, the lack of connection between E and H might suggest that they can be sources of different information even though they are redundant to the same extent. What we need then is *triadic* notion of redundancy and a way of measuring how the information coming to A from E might be *different* from that coming to A from H.

To achieve that we, first, in Section 10.2, develop qualitatively the implications of Burtian theory of structural holes for node-, tie- and triad-level properties of information passing through personal networks of actors. Second, we formalize the process of information flow as a stochastic process in Section 10.3. Third, in order to study these properties we designed a computer simulation described in Section 10.4. Among the results presented in Section 10.5 are five conjectures relating, on the one hand, local properties of the information flow and structural characteristics of the ego-network on the other hand. In particular, the developed a triad-level implication of the theory of structural holes together with the simulation results provide a stronger theoretical justification for the index of *pairwise redundancy* proposed by Bojanowski and Czerniawska-Szejda (2018). We conclude the paper with Section 10.6 by discussing, among other things, what empirical strategies the presented results suggest that are applicable in research settings in which there are multiple types of resources/information being shared through a social network.

10.2 Brokerage and information flow

Actors facing a network information diffusion process can have interests in various aspects of such a situation. Important instance that we will focus on is access to

2 We might expect, as argued by Bojanowski and Czerniawska-Szejda (2018), that structurally non-redundant collaboration ties will be associated with resources that a scientist is unable to acquire through his other collaborations.

information, that is staying informed. An actor might be interested in maximizing chances that any information appearing somewhere in the network will eventually reach him. Other examples include being informed as soon as possible (Buskens and Yamaguchi 1999; Buskens 2002, chap. 4) or maintaining power defined as control over information flow (Reagans and Zuckerman 2008; Emerson 1962). If staying informed is assumed to be a goal motivating tie-forming and tie-dissolving actions of the actors the question of what kind of social networks will emerge endogenously via actors' relational choices becomes a micro-macro problem in the sense of Raub, Buskens, and Van Assen (2011). An example is depicted in Figure 10.2. Macro conditions include assumed properties of the process of information spreading in the network. Properties such as: whether actor loses information when it is shared or not, how the probability of some actor *i* passing information to some actor *j* is defined. Micro conditions include assumptions about the goals and constraints of the actors such as willingness of staying informed while maintaining as few connections as possible. Should the theory of structural holes apply, actors will be likely to drop structurally redundant ties as they are a liability. This would be a micro-level outcome. Removal of redundant ties by the actors leads to appearance of brokers on the macro level. Further, Buskens and Van de Rijt (2008) shown that the value of brokerage positions lies in a "social opportunity structure" that is exhaustible.

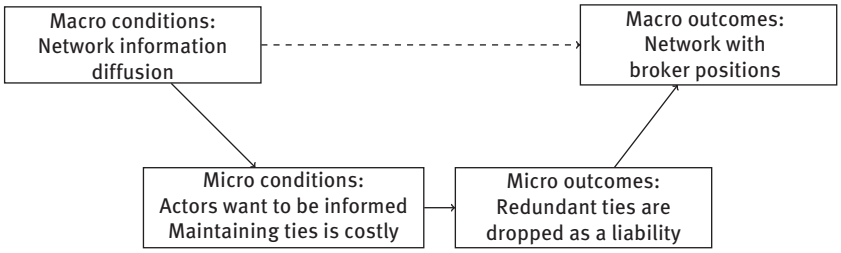


Figure 10.2: Micro-macro mechanism of emergence of brokers when facing information diffusion.

Empirically, when analyzing a single snapshot of a network we usually observe a mixture of broker and non-broker positions. This can be because of the network being far from the structural hole-less equilibrium derived by Buskens and Van de Rijt (2008) or there might be other network formation incentives operating apart from brokerage.

As signaled in the Section 1, brokerage and its properties can be analyzed on different levels: nodes, ties, and triads:

- **Nodes:** Position of a broker in an information flow can be expected to be "information-efficient". That is, every tie should serve a purpose in the sense of being a channel for specific information unavailable through other ties. Statistically

speaking the information type arriving to an actor with high brokerage should be correlated with (predictable from) the alter that it is arriving from.

- **Ties:** Information coming to an actor through a non-redundant tie should be different from that coming through other ties. In statistical terms the distribution of information type coming through a non-redundant tie should be different from the distribution of information type arriving through all the other ties combined.
- **Triads:** Not all redundant ties are alike, even if they are redundant to the same extent. As exemplified with actors *E* and *H* in Figure 10.1, an alter can be more or less redundant to ego *vis a vis* some other alter. In other words, it is a triadic property. We can expect that pairs of alters who are, for example, disconnected in ego's neighborhood to be sources of more different information than pairs of alters who are connected. In statistical terms the distributions of information type in those two pairs will be more similar for connected than for disconnected pairs of alters.

Concrete statistical tools to quantify above mentioned concepts such as brokerage, dyadic and triadic redundancy, as well as for measuring dependence and (dis)similarity of distributions will be introduced in detail in Section 10.5 below.

10.3 The model of information diffusion

Let us have an undirected network $G = \{V, E\}$. Every actor $s \in V$ is endowed with a unique piece of information which will be spreading through the network. At any given time every actor $i \in V$ can be either *informed* or *uninformed* about information *s*. Actor *i* is informed about *s* if the information that started spreading from actor *s* has reached him in the past. He is uninformed about *s* otherwise. As every actor is an origin of unique information his identity (*s*) defines the "information type" – if some actor *i* learns the information that started spreading from some other actor *s* we might say that actor *i* learned information type *s*. The overall history of information status of all actors can be described with an array $M = [m_{ist}]$ such that an element $m_{ist} = 1$ if at time *t* actor *i* is informed about information *s* and $m_{ist} = 0$ otherwise.

In principle the diffusion processes of different types of information may interact. For example, consider some actor *i* informed about information types *s* and *s*'. The probabilities of sharing *s* and *s*' with some neighbor *j* of *i* may not be independent. Actor *i* might share *s* or *s*' at a given time but not both. In this paper we restrict our attention to models in which the processes of diffusion of different types of information are independent. That is, the probability that actor *i* shares information *s* with actor *j* does not depend on whether *i* is informed or uninformed about some other information type *s*', whether he spread *s*' to *j* at the same time step or in the past, etc.

Information spreads through network ties at discrete times. This can be modeled in several ways. For example, Buskens and Yamaguchi (1999) consider two scenarios:

1. the multinomial version in which such information-passing events are independent between actors. The two actors can spread information each to a single neighbor at a single time step
2. the product-binomial version in which information-sharing events are additionally independent within actors. A single actor can spread the information to multiple neighbors (see also Buskens 2002, chap. 4).

Similarly, Borgatti (2005) considers a "gossip model" which is similar to scenario (1) above, but with additional restrictions. First, at each time step the information is shared once. Second, an actor will not share the information with a neighbor who already knows it. There is also a subtle difference in the way Borgatti (2005) implemented the model in their simulations, which we will comment upon in Section 10.4.

The model we analyze in this paper is closest to the Gossip model of Borgatti (2005). We assume that at discrete times an informed actor having at least one uninformed neighbor is selected randomly. She then informs a randomly selected uninformed neighbor. Such processes for each information type s take place independently from each other. Following Buskens and Yamaguchi (1999) and Buskens (2002, chap. 4) for the sake of comparability we represent the model as a finite Markov chain.

Let A_{ij} be the adjacency matrix of graph G and the set of actors who are currently informed as S_1, itself a subset of V. Let us consider the probability that in the next step certain actor u will be informed by somebody from S_1. In other words that in the next time step the set of informed actors becomes S_2 such that $S_1 \subset S_2$ and $S2 \backslash S1 = U$ where $U = \{u\}$. Define

$$r_{ij} = \frac{A_{ij}}{\sum_{j \in V \backslash S_1} A_{ij}}$$

which is a partially-normalized form of the graph adjacency matrix A. Of our interest will be only a submatrix of r_{ij} for $i \in S_1$ and $j \in V \backslash S_1$. Elements of that submatrix are conditional probabilities that uninformed actor j will be informed by i given that i is to inform anybody. Additionally, let $r_{i+} = \sum_{j \in V \backslash S_1} r_{ij}$ are row sums of that submatrix and n_i is the number of elements of r_{i+} which are positive. With that we can define the transition probability of the Markov chain as

$$P(S_1 \rightarrow S_2) = \begin{cases} 0 & \text{if } S_2 \notin S_1 \text{ or } |U| \neq 1 \\ \sum\limits_{i:\, i \in S_1 \wedge r_{i+} > 0} \frac{r_{iu}}{n_i} & \text{otherwise} \end{cases}$$

The chain is defined over the power set of nodes of graph G.

Consider an example network from Figure 10.3. Calculating all transition probabilities leads to a process shown in Figure 10.4. It is worth noting two facts about this stochastic process:

1. There is a single absorbing state corresponding to the situation in which all actors are informed, namely $\{1, 2, 3, 4\}$.
2. All other states constitute singleton transient sets each. In words, each of these states is visited at most once.

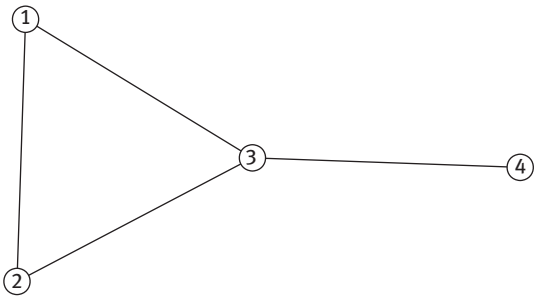


Figure 10.3: Example graph.

To compare alters with respect to the information they bring to ego we need to assess how frequently ego receives particular type of information through a particular alter. We are interested in conditional probabilities of actor i informing actor j with information of type s conditional on actor j being informed about s. In other words, what proportion of times actor j becomes informed about s through neighbor i but not any other neighbor. It does not seem possible to derive such probabilities from the Markov chain defined above. One of the reasons for that difficulty is that defining the state of the chain as the set of informed actors does not provide enough detail on the dynamics of the process which would allow for calculating above mentioned conditional probabilities. To illustrate this, consider the process in the state $\{1, 3\}$ and transitioning to the state $\{1, 2, 3\}$. In such transition actor 2 is getting informed actor 1 or actor 3. We know from above calculation that the transition probability is 0.75 – given that actors 1 and 3 are informed in the next step actor 2 will be informed (with probability 0.75) or actor 4 will be informed (with probability 0.25). What we do not know is whether actor 2 learned the information from actor 1 or from actor 3. The conditional probabilities of these two events are not equal:

- Given that actor 1 is to inform anybody, she will inform actor 2 with probability 1
- Given that actor 3 is to inform anybody, she will inform actors 2 and 4 each with probability 0.5

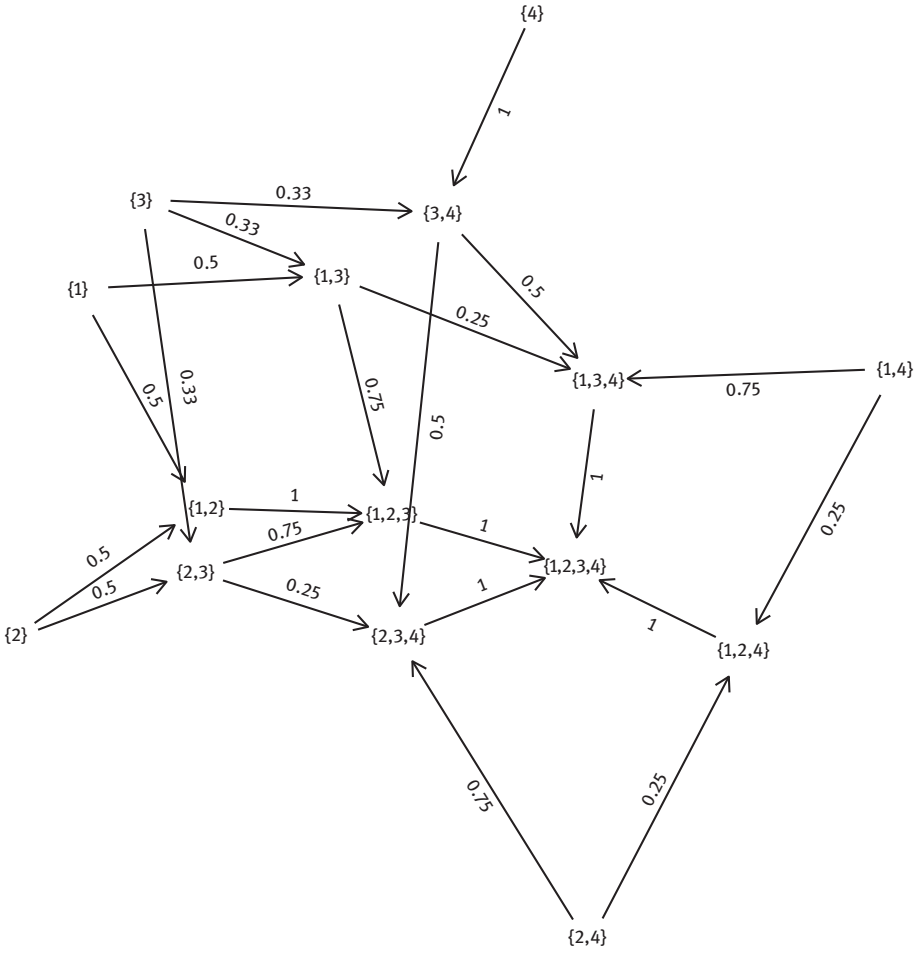


Figure 10.4: Markov Chain of information diffusion over the example graph. Each state (graph node) is a subset of actors who are informed. Arcs of the graph are state transitions labelled with transition probabilities. The set $\{1, 2, 3, 4\}$ is an absorbing state.

Assuming that a random informed actor is to inform somebody, we select one actor from actors 1 and 3 at random. It then follows that:

- Actor 3 will inform actor 4 with probability $0.5 \cdot 0.5 = 0.25$.
- Actor 3 will inform actor 2 with probability $0.5 \cdot 0.5 = 0.25$.
- Actor 1 will inform actor 1 with probability 0.5.

Such calculations of conditional probabilities become more complex in a generic case and on a larger graph. We therefore turn to a numerical simulation to approximate them. This is presented in the next section.

10.4 Simulation study

We designed a computational experiment in which a single run consists of the following steps:

1. Take an undirected connected graph $G = \{V, E\}$.
2. Define a node attribute partitioning nodes into informed and uninformed. Initially all nodes are uninformed.
3. Select a node, called *seed*, and set him as informed.
4. Time progresses in discrete steps. In every time step the information is shared by a random informed node with a random network neighbor who is uninformed. In particular:
   a) Identify ties that connect an informed node and an uninformed node
   b) Identify informed nodes incident on the ties listed in (a)
   c) Select a random node from those found in (b)
   d) The informed node identified in (c) shares the information with a random uninformed neighbor.
5. Repeat (4) until all the nodes are informed.

The fact the network is connected guarantees that every node will ultimately learn the information. The design of the information diffusion model implemented in step 4 above additionally guarantees that there are no "idle" simulation steps as on every step some informed node will share the information with an uninformed node. In consequence, each simulation run always consists of exactly $|V| - 1$ steps. The process is almost identical to the "gossip process" used by Borgatti (2005). The only difference is in the stopping criterion (item 5 above). Primarily for the purpose of comparability with other types of network flow processes Borgatti (2005) used a setup in which a random seed and "target" nodes were selected. The simulation was complete when starting from a single informed "source" node the process reached the state in which the "target" node became informed.

For the purpose of this paper we use the Zachary's Karate Club network (Zachary 1977). It consists of 34 nodes and 78 ties presented in Figure 10.5. We run[3] the simulation 100 times for every seed node amounting to the total of $34 \times 100 \times 33 = 112200$ information-sharing events.

The data generated by the simulation consists of a set of information sharing events each being described with the following variables:

- B – simulation run with values from $\{1, \ldots, 100\}$
- S – ID of the seed actor with values from $V = \{1, \ldots, 34\}$
- I – ID of the informed actor (information sender) with values from $V = \{1, \ldots, 34\}$

3 The simulation was implemented using R (R Core Team 2019) and packages: "igraph" (Csardi and Nepusz 2006), "netflow" (Bojanowski 2019), "graphlayouts" (Schoch 2019).

- J – ID of the uninformed actor (information receiver) with values from $V = \{1, \ldots, 34\}$
- T – Time step at at which actor I shared information S with actor J. A number from $\{1, \ldots, 33\}$.

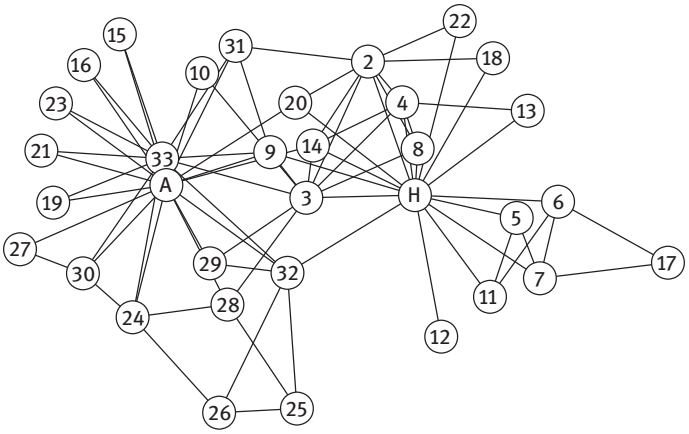


Figure 10.5: Zachary's Karate Club network used in simulations.

In words, every row is an event that took place in simulation run B at time T in which the information that started spreading from seed actor S was passed by actor I to actor J.

10.5 Simulation results

To analyze the simulated data, we are interested in studying statistical relationships between structural characteristics of personal networks on the one hand and local properties of the information flow on the other. Following the arguments from Sections 10.1 and 10.2 we discuss the results in the following subsections split by the level of analysis: actors, ties (dyads), and triads respectively.

10.5.1 Actors

Our intuitions from Section 10.2 were that personal networks of actors in brokerage positions will be characterized with a kind of "information efficiency" in which the message type is highly correlated with the identity of the alter through whom the actor learns it. To quantify this intuition we need measures for both concepts. Popular

ways of capturing the extent of brokerage in a personal network is *Constraint* (Burt 1995, 54) defined as

$$c_i \equiv \sum_{j \neq i} (A_{ij} + \sum_{k \neq i,\, k \neq j} A_{ik} A_{kj})^2$$

where A is, as above, the adjacency matrix of the graph. Another one is *network efficiency* (Borgatti 1997) defined as:

$$e_i \equiv \frac{1}{n}\left(n - \frac{2t}{n}\right) = 1 - \frac{2t}{n^2}$$

where t is the number of alter-alter ties in ego's neighborhood and n is the number of alters.

Assessing the dependence between message type and alter identity requires an index of dependence between variables S (information type) and I (alter ID) calculated separately for each value of J (information receiver) – the ego. As the variables S and I are nominal, we need an index of *stochastic dependence* for which we choose the information-theoretic quantity (Lissowski 1977):

$$\kappa \equiv \frac{H(S) - E(H(S|J))}{H(S)}$$

where $H(\cdot)$ is the entropy of a, possibly conditional, distribution of a variable. The index varies between 0 (stochastic independence) and 1 (functional dependence), and has a convenient interpretation as a proportional reduction in prediction error (e.g. Agresti and Finlay 1997, chap. 8.7).

Figures 10.6 and 10.7 show actors in the Karate network plotted according to the strength of dependence between message type and alter ID (vertical axis) against two measures of brokerage (horizontal axis): Burt's constraint and network efficiency. In Figure 10.6 we see that actors with low values of constraint, the brokers, are characterized with higher dependence and the higher the constraint, the weaker the dependence. In Figure 10.7 the higher the network efficiency (indicating higher brokerage) the stronger the dependence. This leads us to:

Conjecture 1: (Brokers) In social networks actors in brokerage positions are characterized with a relatively strong dependence between (1) type of information arriving (identity of the seed node) and (2) an alter that is the source of the information for the actor.

10.5.2 Ties

We need to quantify two concepts: (1) the redundancy of alter j to ego i and (2) the extent to which the information coming from alter j to ego i is different from

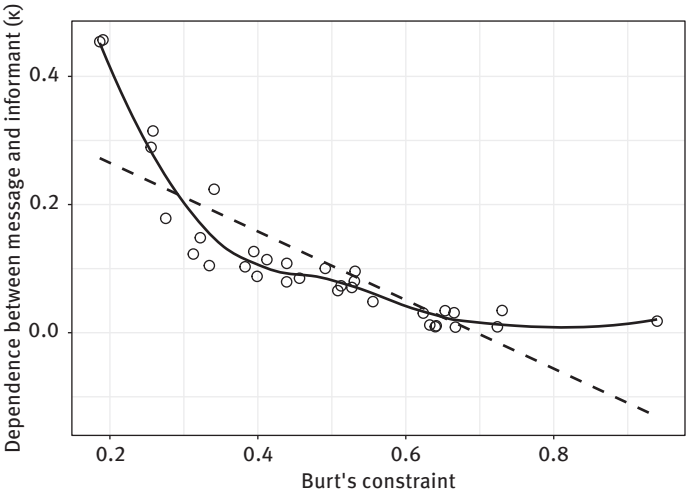


Figure 10.6: Strength of dependence between message type and informant versus Burt's constraint. For brokers (low constraint) it is more predictable what information comes from which alter. Linear (dashed) and LOESS (solid) trends superimposed.

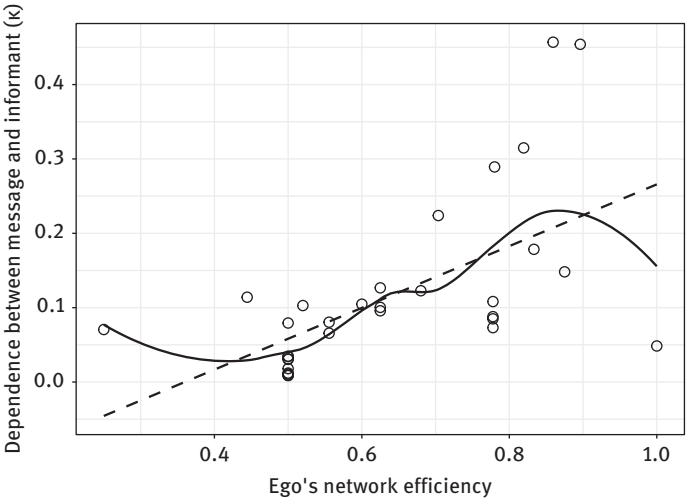


Figure 10.7: Strength of dependence between message type and informant versus ego's network efficiency. For brokers (high ego-network efficiency) it is more predictable what information comes from which alter. Linear (dashed) and LOESS (solid) trends superimposed.

information coming to ego from the remaining alters. Common way of measuring alter's redundancy is *dyadic redundancy* (Hanneman and Riddle 2005, chap. 9):

$$\text{dyadic redundancy}_{ij} \equiv \frac{d_{ij}}{n_i}$$

where d_{ij} is number of alter-alter ties in the neighborhood of ego i adjacent to actor j and n_i is the number of alters of ego i.

To compare different distributions we will use Kullback-Leibler Divergence (Kullback and Leibler 1951; Kullback 1959). For two discrete probability distributions $p_i = P(X = x_i)$ and $q_i = P(Y = y_i)$ of some variables X and Y it is equal to

$$\mathrm{KLD}(X, Y) = \sum_i p_i \log \frac{q_i}{p_i}$$

It is equal to 0 if and only if distributions of X and Y are identical. In order to compare the distribution of information coming to ego from an alter to the distribution of information coming to ego from all other alters we are interested in calculating $\mathrm{KLD}(S|J = j \wedge I = i,\ S|J \neq j \wedge I = i)$.

Figure 10.8 shows all ego-alter ties in the Karate network according to dyadic redundancy (horizontal axis) and KL divergence calculated as above. We can see that alters characterized with low redundancy are more likely to bring different information to ego than those with high redundancy. This can be summarized as:

Conjecture 2: (Redundant ties) In personal social networks the two distributions of information types coming to an actor (1) from an alter and (2) coming from the remaining alters will be more different the lower the dyadic redundancy of that alter.

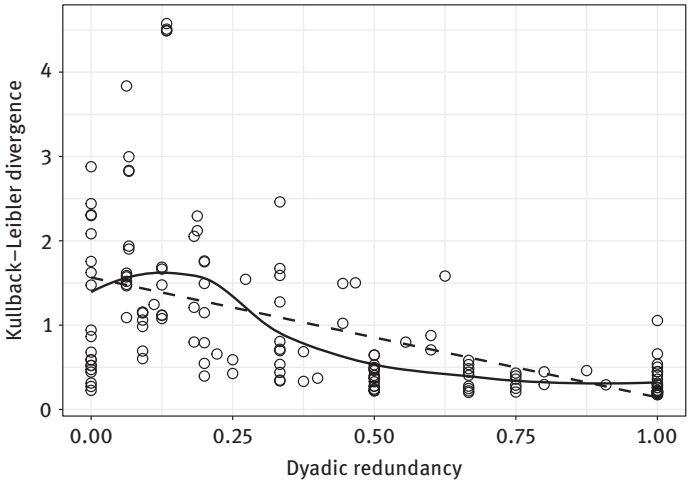


Figure 10.8: Dyad information divergence versus dyadic redundancy. Non-redundant ties (low dyadic redundancy) tend to bring different information than other ties. Linear (dashed) and LOESS (solid) trends superimposed.

Thus both conjectures are along the initial intuitions. Firstly, on the actor level, that brokerage positions are information-efficient. Secondly, on the tie level, that non-redundant ties tend to bring different information than other ties.

10.5.3 Triads

We have argued in Section 10.2 that the consequences of occupying brokerage positions in a social network and information diffusion extend beyond the actor and dyadic levels. To make these intuitions more concrete we are going to analyze the simulated data on a triadic level by focusing on triples of actors consisting of ego *j* and two alters *i* and *k* and the following properties:

1. The (dis)similarity between (a) distribution of information type that arrives to ego *j* from alter *i* and (b) distribution of information type that arrives to ego *j* from alter *k*.
2. A measure capturing the extent of redundancy of alters *i* and *k* to ego *j*.

We will measure the dissimilarity between the distributions (item 1 above) with Kullback-Leibler Divergence as defined in the previous section. To capture the triadic notion of redundancy (item 2) we may look at geodesic distance between alters *i* and *k* in ego's neighborhood. The closer the alters are the more they are redundant. Figure 10.9 shows all ego-alter-alter triplets for which the alter-alter distance is finite according to the alter-alter distance (horizontal axis) and KL divergence measuring dissimilarity between information type distributions for the two alters in the triplet. We can indeed observe that the further the two alters are apart in the triplet the more dissimilar are the distributions of information type that arrive from these alters to ego. We summarize this with

Conjecture 3: (Alter-alter distance) In personal social networks the distributions of information types coming to the ego from two alters are the more dissimilar the greater the distance between these alters in ego's neighborhood.

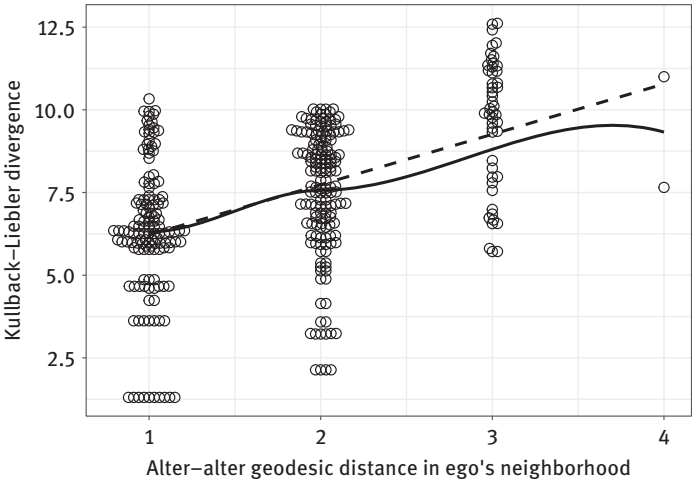


Figure 10.9: In ego-alter-alter triads in which alters are further apart, alters tend to bring more different information. Linear (dashed) and LOESS (solid) trends superimposed.

Using alter-alter distance to indicate the extent of redundancy has two limitations. The first one is that we do not account for the pairs of alters who are disconnected. Distance between them is by convention equal to infinity. The second one is that we treat pairs of directly connected alters (at distance 1) as redundant to the same extent irrespectively of their embeddedness in the personal network of ego. To address the second issue we may capture that embeddedness by counting how many alters the pair of alters under study has in common. In the example network in Figure 10.1 actors *G* and *H* have no other alters in common while the pair *B*-*D* has one alter in common – actor *C*. We might expect the more embedded a pair of alters is the more similar the distributions of information types will be. This is what we observe in Figure 10.10 which shows all ego-alter-alter triples in the Karate network according to the number of shared alters of the alter-alter pair (horizontal axis) and KL divergence of the distributions of information type coming from the two alters to ego. The more embedded the alter-alter pair is the less dissimilar the distributions. Hence the conjecture

Conjecture 4: (Alter-alter embeddedness) In personal social networks the distributions of information type coming to the ego from two alters are the less dissimilar the more embedded the alter-alter pair is in ego's neighborhood.

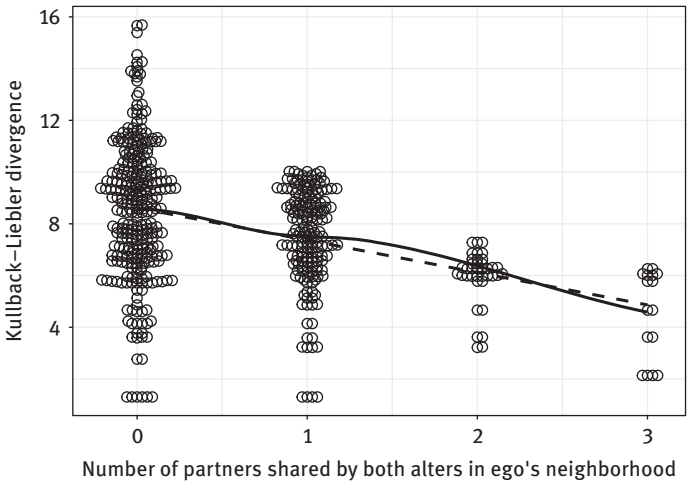


Figure 10.10: In ego-alter-alter triads alters sharing more partners tend to bring similar information. Linear (dashed) and LOESS (solid) trends superimposed.

Table 10.1 shows a cross-classification of all ego-alter-alter triples in the Karate network. While number of shared partners (in columns) is shown as is, we have transformed the alter-alter distance (in rows) by (1) showing the inverse and (2) assigning value 0 to those triples in which the alters are not connected directly or indirectly.

Notice that alter-alter distance and alter-alter embeddedness are almost complementary characteristics describing the neighborhood of of an ego. The (inverse) distance between the alters takes non-zero values only for those ego-alter-alter triples in which there are no shared partners indicating a null alter-alter embeddedness. This suggests that the two characteristics can be combined into a single numerical index.

Table 10.1: Cross-tabulation of inverse alter-alter distance and number of shared alters. The two measures are almost complementary.

Inverse alter-alter distance	Number of shared alters			
	0	1	2	3
0.0000000	182	0	0	0
0.2500000	2	0	0	0
0.3333333	49	0	0	0
0.5000000	0	154	2	4
1.0000000	63	24	36	12

Let $d(j,i,k)^{-1}$ be the inverse of the length of the shortest path between alters i and k in personal network of actor j having all of j's ties removed. By convention we assume $d(j,i,k)^{-1}=0$ if the path i and k does not exist. Further, let $\mathrm{sh}(j,i,k)$ be the number of alters of ego j who are directly connected to alters i and k. The combined index, which we call *pairwise redundancy* (Bojanowski and Czerniawska-Szejda 2018) can be defined as:

$$\mathrm{PR}(j,i,k)=\begin{cases}0 & \text{if } d(j,i,k)=0 \text{ and } \mathrm{sh}(j,i,k)=0\\ d(j,i,k)^{-1} & \text{if } d(j,i,k)>1 \text{ and } \mathrm{sh}(j,i,k)=0\\ 1 & \text{if } d(j,i,k)=1 \text{ and } \mathrm{sh}(j,i,k)=0\\ \mathrm{sh}(j,i,k)+1 & \text{if } d(j,i,k)=1 \text{ and } \mathrm{sh}(j,i,k)>0\end{cases}$$

In words:

- It is 0 if alters i and k of ego j are not connected directly or indirectly, e.g. alter F vs all other alters of ego A in Figure 10.1.
- It is in the interval (0, 1) if alters i and k of ego j are connected only indirectly, e.g. alters B and E of ego A in Figure 10.1. It is the inverse of the shortest path between j and k.
- It is 1 if alters i and k are connected directly with no shared partners, e.g. alters G and H of ego A in Figure 10.1.
- It is in the interval $(1,\infty)$ if alters i and k are connected directly and have common some of ego's alters in common. It is the number of shared partners plus 1. For example for alters B and D in Figure 10.1 it is $\mathrm{PR}(A,B,D)=1+1=2$ as alters B and D have alter C in common.

For illustration Table 10.2 contains all the pairwise redundancy scores for alters of ego *A* in the network from Figure 10.1.

Table 10.2: Pairwise redundancy scores for pairs of alters of actor A in the network from Figure 10.1.

Alter 1	Alter 2					
	C	D	E	F	G	H
B	2	2	0.5	0	0	0
C		2	0.5	0	0	0
D			1	0	0	0
E				0	0	0
F					0	0
G						1

We can now use the pairwise redundancy scores to characterize all ego-alter-alter triples from the Karate network jointly and set it against the KL divergence scores comparing the distributions of information type arriving from the alters to ego. This is shown in Figure 10.11. We can see that the more pairwise-redundant the alters are, the more similar the distributions of information type coming from these alters:

Conjecture 5: (Pairwise redundancy) In personal social networks the distributions of information type coming to the ego from two alters are the less dissimilar the more pairwise-redundant the alters are to ego.

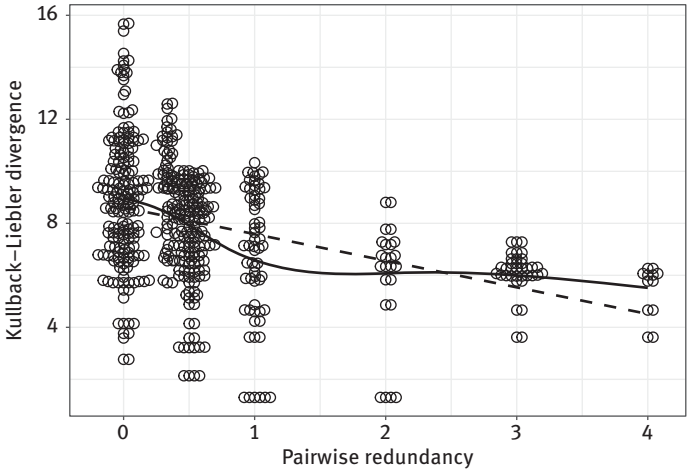


Figure 10.11: Pairwise-redundant alters tend to bring similar information. Linear (dashed) and LOESS (solid) trends superimposed.

Conjectures 3, 4, and 5 formulated in this section are along the qualitative arguments discussed earlier, namely that there is a triadic dimension to brokerage and redundancy. Alters who might be redundant to the same extent may be at the same time non-redundant vis a vis each other in terms of the divergence of incoming information.

10.6 Conclusion and discussion

In the presented paper we have formalized a model of information diffusion and used computer simulation to show how statistical properties of information flow are related to brokerage characteristics on different levels of analysis. The results encapsulated in conjectures 1–5 provide us with viable empirical strategies for answering research questions regarding possible impact of the network structure on the tie content, if the latter is understood as the diversity of the types of information/resources being transferred or shared along that tie.

Consider a research setting in which there are multiple types of resources or information spreading in the network. Collected data could have a form of a set of dyadic variables, one for each resource, combined into an array, say, $R = [r_{ijk}]$. Elements of R could be binary, e.g.: $r_{ijk} = 1$ if resource k has been shared/passed from actor i to actor j and $r_{ijk} = 0$ otherwise. Alternatively r_{ijk}'s could be frequencies of how often actor j used resource k provided by i in some period of time. Symptoms of brokerage could be then analyzed on the three different levels we covered in the conjectures, namely:

- **Node level:** By Conjecture 1 we expect stronger dependence of information type and alter for brokers than for non-brokers. It is then sensible to characterize every actor j with the strength of statistical association in the corresponding layers $r_{i(j)k}$. The actual choice of the measure of association depends on the measurement level of r_{ijk}'s.
- **Tie level:** By Conjecture 2 we expect the non-redundant ties to be characterized with higher dissimilarity between distribution of incoming information types and the distribution of information incoming from other alters. Consequently, we may compare the dissimilarity of a bundle acquired by actor j from actor i described by the vector $r_{(ij)k}$ to aggregated bundles from all other neighbors of j.
- **Triad level:** By Conjectures 3, 4, and 5 we may compare resource bundles acquired by actor j from actor i to a resource bundle acquired by actor j from some other actor i'. If r_{ijk}'s are binary a reasonable choice for dissimilarity measure is the Jaccard coeficient. The ego-alter-alter triads characterized with higher *pairwise redundancy* are then expected to be characterized with higher dissimilarity. An approach of this kind was used by Bojanowski and Czerniawska-Szejda (2018).

A critical reader may notice that associations behind the formulated conjectures illustrated in Figures 10.6 through 11 show quite some variation. It is to be expected as our analyzes are based on personal networks of the actors. The choice was intentional as egocentrically sampled data is popular empirical material in the social sciences. The variation, at least in part, comes from higher-order structural properties of the network. In particular, higher-order redundancies. A natural follow-up research question could therefore be whether and how a concept similar to pairwise redundancy could be formulated when complete network data is available. An intuition already formulated by Burt (1995) points to ideas of "redundancy by equivalence".

More immediate problems are related to strengthening the presented results. Firstly, it should be possible to derive the conditional diffusion probabilities analytically. It seems promising to represent the information diffusion process formulated as a Markov chain in Section 10.3 terms of diffusion trees. The results are elusive though. Secondly, it is interesting whether the presented results will still hold under different variants of the information diffusion model. Obvious alternatives are the multinomial and product-binomial models of Buskens and Yamaguchi (1999). It seems that in the case of the multinomial model the differences should not be substantial. We leave these questions for further research.

References

Agresti, Alan, and Barbara Finlay. 1997. *Statistical Methods for the Social Sciences*. 3rd ed. New Jersey: Prentice Hall.

Bojanowski, Michal. 2019. *Netflow: Simulating Network Flows* (version of R package 0.0-1). https://github.com/mbojan/netflow.

Bojanowski, Michał, and Dominika Czerniawska-Szejda. 2020. "Reaching for Unique Resources: Structural Holes and Specialization in Scientific Collaboration Networks." *Journal of Social Structure* 21(1): 1–34. https://doi.org/10.21307/joss-2020-001.

Borgatti, Stephen P. 2005. "Centrality and Network Flow." *Social Networks* 27 (1): 55–71.

Borgatti, Stephen P. 1997. "Structural Holes: Unpacking Burt's Redundancy Measures." *Connections* 20 (1): 35–38.

Burt, Ronald S. 1995. *Structural Holes: The Social Structure of Competition*. Cambridge, MA: Harvard University Press.

Burt, Ronald S., Robin M. Hogarth, and Claude Michaud. 2000. "The Social Capital of French and American Managers." *Organizational Science* 11 (2): 123–47. https://doi.org/10.1287/orsc.11.2.123.12506.

Buskens, Vincent. 2002. *Social Networks and Trust*. Dordrecht: Kluwer.

Buskens, Vincent and Arnout van de Rijt. 2008. "Dynamics of Networks If Everyone Strives for Structural Holes." *American Journal of Sociology* 114 (2): 371–407.

Buskens, Vincent, and Kazuo Yamaguchi. 1999. "A New Model for Information Diffusion in Heterogeneous Social Networks." *Sociological Methodology* 29 (1): 281–325.

Csardi, Gabor, and Tamas Nepusz. 2006. "The Igraph Software Package for Complex Network Research." *InterJournal* Complex Systems: 1695. http://igraph.org.

Emerson, Richard M. 1962. “Power-Dependence Relations.” *American Sociological Review*, 27(1): 31–41.

Granovetter, Mark. 1974. *Getting a Job: A Study of Contacts and Careers*. Chicago: University of Chicago Press.

Granovetter, Mark S. 1977. “The Strength of Weak Ties.” In *Social Networks: A Developing Paradigm*, edited by Samuel Leinhardt, 347–67. New York: Academic Press.

Hanneman, Robert A., and Mark Riddle. 2005. *Introduction to Social Network Methods*. University of California, Riverside. http://faculty.ucr.edu/~hanneman/.

Kullback, Solomon. 1959. *Information Theory and Statistics*. New York: John Wiley & Sons.

Kullback, Solomon, and Richard Leibler. 1951. “On Information and Sufficiency.” *Annals of Mathematical Statistics* 22 (1): 79–86.

Lissowski, Grzegorz. 1977. “Statistical Association and Prediction.” In *Problems of Formalization in the Social Sciences*, edited by Klemens Szaniawski, 217–45. Wrocław: Ossolineum.

Raub, Werner, Vincent Buskens, and Marcel Van Assen. 2011. “Micro-Macro Links and Microfoundations in Sociology.” *The Journal of Mathematical Sociology* 35 (1–3): 1–25.

R Core Team. 2019. *R: A Language and Environment for Statistical Computing* (version 3.5.2). Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Reagans, Ray, and Ezra Zuckerman. 2008. “Why Knowledge Does Not Equal Power: The Network Redundancy Trade-Off.” *Industrial and Corporate Change* 17 (5): 903–44.

Schoch, David. 2019. *Graphlayouts: Additional Layout Algorithms for Network Visualizations* (version of R package 0.0.5.9000). https://github.com/schochastics/graphlayouts.

Zachary, Wayne W. 1977. “An Information Flow Model for Conflict and Fission in Small Groups.” *Journal of Anthropological Research* 33 (4): 452–73.

Thomas Gautschi

11 Who Gets How Much in Which Relation? A Flexible Theory of Profit Splits in Networks and its Application to Complex Structures

Abstract: Starting from exogenously given negotiation networks, sociological exchange theories explain bilateral divisions of fixed surpluses (e.g., cake, dollar) as consequences of the respective partners' structural embeddedness. There are many competing theories, most of which are, despite of their differences, exclusively concerned with explaining outcomes in pure negatively connected networks with a one-exchange rule. That is, the modelling efforts are directed towards scenarios in which an actual exchange in one relation prevents transfers in other relations and, in addition, only one exchange per round is allowed. Such a narrow focus seems unnecessary since experimental results for networks with more complex relational characteristics are available: (i) networks with positively connected negotiation ties (Yamagishi, Gillmore, and Cook 1988), (ii) networks with varying cake sizes to be partitioned (Bonacich and Friedkin 1998), and (iii) networks where positions differ with respect to the number of exchanges they need to complete (Lovaglia et al. 1995; Skvoretz and Willer 1993; Willer and Skvoretz 1999). In this paper we present a simple model which is not limited to a particular class of exchange networks but allows for point predictions in networks with (all combinations of) complex characteristics. Specifically, we combine the generalized Nash bargaining solution from cooperative game theory with the assumption that both relational features and network positions affect exchange outcomes. Its predictions correspond closely with experimental results.

11.1 Introduction

Sociological exchange theories seek to explain the bilateral division of a fixed and perfectly divisible surplus (e.g., cake, dollar). Such a positive surplus may, for instance, materialize if the buyer's willingness to pay exceeds the seller's reservation price. Exchange theories then assume that ac- tors agree on the partitioning of the surplus in bilateral negotiations. And, exchange theorists agree with Emerson (1972, 1981) that negotiations and exchange rarely occur in isolate encounters. They thus consider situations in which a given network structure defines matches between pairs of bargaining partners (Cook et al. 1983) and attribute exchange outcomes in dyads to the effects structure has on individual negotiation success. Exchange theory

Thomas Gautschi, University of Mannheim

https://doi.org/10.1515/9783110647495-011

thus purposely neglects effects other than structure on exchange outcomes (e.g., effects of individual characteristics such as age, sex or education).

In accordance with this assumption, the basic premise of sociological exchange theory is that actors seek to benefit from exchange by exploiting structural advantages. That is, they use their network positions to maximize their individual shares of the cakes to be negotiated on in dyadic bargaining. Theorists thus prefer a rational actor perspective in the sense that interactive choices are characterized by purposive and profit maximizing individual behavior. Consequently, dyadic exchange will produce outcomes in which one actor will not necessarily obtain as much exchange profit as the other. And, exogenous restrictions may force actors to choose an exchange partner from a larger set of negotiation partners. In accordance with these ideas, sociological exchange theory thus searches for answers on two fundamental questions. First, who gets, depending on his position in the negotiation structure, how much of the surplus? And second, who exchanges with whom if, after negotiations with different partners, an exchange partner has to be chosen? Put differently, can structural deviations between the bargaining and exchange networks be identified?[1] Besides the consensus to study the effects of structure only on exchange outcomes, sociological exchange theories have other common features. At least partly, these features reflect that theory building was closely related with the restricted exchange situations studied in laboratory experiments. Most experimental research focuses on the study of exchange structures with negatively connected and equally valued relations.[2] The latter refers to exchange systems where relations concern the split of an identical surplus. And, negatively connected systems are characterized by the fact that exchange in one relation tends to prevent exchange in other relations (e.g., dating networks). Accordingly, the close relationship between model building and experimental work lead to the situation that the focus was put on developing theories for negatively connected exchange structures only.[3] Even though most

1 While we do not address the latter question in this paper either, we show in Braun and Gautschi (2007) how our model (or any other model which provides point predictions on the division of cakes) can be used to decide whether given bargaining ties become exchange patterns if there is competition between potential trading partners. However, there are publications which endogenize the formation, stabilization, and modification of actor relations in networks (e.g., Aumann and Myerson 1988; Bala and Goyal 2000; Demange and Wooders 2005; Dutta and Jackson 2003; Gould 2002; Jackson and Wolinsky 1996; Skyrms and Pemantle 2000; Slikker and van den Nouweland 2001; Vega-Redondo 2007; Watts 2001) and thus treat the network as the "explanatory variable".

2 Exceptions are Yamagishi, Gillmore, and Cook's (1988) experiments on a positively connected 5-LINE and Bonacich and Friedkin's (1998) experiments on simple exchange structures with variations in the value of the cake to be partitioned. Networks with positively connected exchange ties are normally associated with situations in which resource transfers in one relation tend to promote transfers in other relations (e.g., communication networks).

3 Modifying and extending Coleman's (1973, 1990) competitive equilibrium approach, Yamaguchi (1996) is the only theory which can be applied to negatively and positively connected exchange

established theories can be adjusted to account for variations in cake sizes to be divided (e.g., Friedkin's (1986, 1992) Expected Value theory, Skvoretz and Willer's (1993) Exchange Resistance theory, Yamagishi and Cook's (1992) Equi-Dependence theory, and partly, Markovsky, Willer, and Patton's (1988) GPI measure), the focus of most theories nevertheless appears considerably narrow and seems an unnecessary restriction. Especially since Cook et al. (1983), Cook and Emerson (1978), and Emerson (1972) have established the distinction between negatively and positively connected exchange relations quite some time ago.

On the one hand, as claimed above, modeling efforts are confined by its close connection to the experimental focus on negatively connected exchange structures and equally valued relations (i.e., homogeneity with respect to cake sizes). On the other hand, however, basic characteristics of the experimental bargaining protocol are still not reflected by established theories[4] such that they are, at least, partly based on divergent assumptions. In short, two points are most relevant. First, all theories assume that negotiation partners pursue their self-interest, and such is stimulated in laboratory experiments. However, the theories usually do not specify an optimization problem. And, it is unclear where network features enter the choice calculus and how they influence decision making. Second, interactive choices characterize (experimental) negotiation processes. However, only a few theories explicitly refer to game-theoretic ideas but none uses a dyadic bargaining solution with a solid foundation in non-cooperative bargaining theory. Not one theory thus reflects the experimentally implemented bargaining protocol of alternating offers. This divergence between experiments and theories seems questionable since experimental results make up the measure rod against which the theories' performance is tested.

structures. He equates negative (positive) connections with closely substitutable (complementary) exchange relations and introduces a flexible continuous parameter for substitutability/complementarity (viz., the elasticity of substitution). Yamaguchi's theory thus embraces situations in which exchange in one relation tends to prevent or promote transfers in others. Its application requires, however, an ad hoc specification of the elasticity of substitution. And, it is limited to the analysis of either substitutable or complementary relations in a given network. Combining basic ideas of his original model with additional assumptions, Yamaguchi (2000) presents a theoretical analysis of structures characterized by the simultaneous presence of both substitutability and complementarity among the multiple exchange relations of an actor.

4 There are a wide range of different theories but about as many controversies among them as well (e.g., Bienenstock and Bonacich 1992, 1993, 1997; Bonacich 1998, 1999; Bonacich and Bienenstock 1995; Bonacich and Friedkin 1998; Burke 1997; Friedkin 1992, 1993, 1995; Lovaglia et al. 1995; Markovsky et al. 1993, 1997; Markovsky, Willer, and Patton 1988, 1990; Skvoretz and Fararo 1992; Skvoretz and Lovaglia 1995; Skvoretz and Willer 1991, 1993; Thye, Lovaglia, and Markovsky 1997; Yamagishi and Cook 1990; Yamagishi, Gillmore, and Cook 1988; Yamaguchi 1996, 1997, 2000). Special issues of journals (cf., *Social Networks* 14, No. 3–4, 1992 and, at least partly, *Rationality and Society* 9, No. 1–2, 1997) contain additional articles and controversies. Willer and Emanuelson (2008) test ten prominent theories on their empirical relevance.

Current sociological exchange theories are limited in scope which restricts them in being able to make predictions for exchange systems with more realistic features. Especially given the empirical accuracy of established theories in predicting outcomes in negatively connected structures, researchers should make the necessary further step, adjusting and revising their theories (or developing new theories) to be able to deal with more complex networks (viz., negatively connected, positively connected and mixed network structures with potentially unequally valued relations and no restrictions on node-specific parameters such as the number of exchanges a position intends to complete). Results from experiments employing networks with more complex relational and positional characteristics are publicly available. Yamagishi, Gillmore, and Cook (1988), for instance, present results for a positively connected 5-Line. Positive connections were ensured by (i) introducing distinct resources at opposite ends of the line structure, (ii) dictating that resource *x* be traded for resource *y* in one relation and *y* for *x* in the other relation, and (iii) by paying profit points only for pairs of resources collected (complementary goods). Moreover, Yamagishi, Gillmore, and Cook correctly emphasize that real networks often are mixtures of both, negatively and positively connected relations. Consequently, they additionally provide experimental results for a mixed network structure.[5] Experimental evidence of yet another kind is offered by Bonacich and Friedkin (1998). They discuss results from an experiment on four network structures, each of which was characterized by unequally valued relations. As their results show, profit distributions are considerably affected, compared to a system of equally valued relations, by the heterogeneity in the surpluses to be divided in bilateral negotiation. Finally, Skvoretz and Willer (1993) and Willer and Skvoretz (1999) discuss experimental findings from networks with equally valued relations but where positions differed with respect to the number of exchanges they needed to complete. These networks were thus characterized by a heterogeneity in the number of exchanges positions conclude per round of negotiation and exchange. For those structures where results from equivalent networks with a one-exchange rule are available, it is found that dyadic profit splits indeed differ from the simple (homogeneous) scenario.

While there are theories which can, at least partly, deal with some of the more complex network characteristics (e.g., Markovsky et al. 1993; Skvoretz and Willer 1993; Yamaguchi 1996, 2000), there is not a single model which could consistently address all aspects of complex networks. Moreover, some theories which can partly handle complex networks are adjusted versions of existing theories which have been revised such that ad hoc assumptions now delineate their updated versions. This indicates the need for a sufficiently general theory of exchange networks.

5 This relational assessment was ensured by 'adding' two new actors to one of the peripheral positions in the positively connected 5-LINE. This position can now obtain identical resources (say, *x*) through either of his new contacts. Since they offer the same resource, they are thus negatively connected to the former peripheral 5-LINE actor.

First of all, such a model should be parsimonious and simple to apply. At the same time, if researchers stress that negotiation partners pursue their self-interest (i.e., the rationality postulate stimulated in laboratory experiments), a theory should clarify how structural features affect the actors' decision-making and how their interactive choices determine the negotiation outcomes. Second, a model may not be limited to the analysis of negatively connected settings but must embrace all relational assessments of ties. Third, it should allow for unequally valued relations (i.e., variations in terms of the surplus to be partitioned) and positional heterogeneity in the number of exchanges positions wish to complete per round. Finally, the model should offer unique point predictions for negotiation outcomes which closely fit the available experimental evidence. That is, it should be able to predict empirical evidence from simple as well as from more complex bargaining experiments.

We present and apply a model with these properties in the remainder of this article. This model combines the generalized Nash (1950, 1953) bargaining solution from cooperative game theory (cf., Binmore 1992) with a specific definition of each actor's bargaining power in terms of relational as well as positional features and network embeddedness.[6]

11.2 Relational assessment

Even though the classification of negatively and positively connected network relations (Cook and Emerson 1978; Cook et al. 1983; Emerson 1972) is widely accepted in the sociological exchange literature, there is no sound and formal definition for negative or positive ties. The lack of such a definition poses no problem in the case of pure negatively or positively connected negotiation structures. However, Yamagishi, Gillmore, and Cook (1988) correctly emphasize that real networks often are mixtures of both types of relations (e.g., exchange in the Kula Ring as described by Malinowski 1922). The loose definition of negatively and positively connected ties – that is, exchange in one relation either prevents or promotes transfers in others – is unfortunately not sufficient for a precise mathematical description of mixed negotiation networks. Consequently, unique profit point predictions for mixed networks are not necessarily guaranteed. From a methodological point of view it is, however, essential

6 Since we first presented this basic idea (Gautschi 2002), other authors also suggested exchange theories based on the Nash bargaining solution (Bayati et al. 2015; Chakraborty and Kearns 2008; Chakraborty et al. 2009; Kleinberg and Tardos 2008). The cooperative bargaining solution coincides with the subgame-perfect Nash equilibrium outcome of Rubinstein's game if the focus is on the limiting scenario in which the amount of time between proposals vanishes (cf., Binmore 1985, 1998; Muthoo 1999; Osborne and Rubinstein 1990). The generalized Nash bargaining solution thus is an appropriate cooperative solution concept because it can be derived from a strategic analysis in the sense of non-cooperative game theory.

to have a definition of negative and positive connection which allows for an unequivocal characterization of network ties. Only this provides a basis for unique profit point predictions in all types of networks.

In experiments, positive connections are ensured by demanding that two goods flow through the network (Yamagishi, Gillmore, and Cook 1988).[7] Since actors can derive utility from pairs of goods only – that is, goods which are always consumed together in fixed proportions (perfect complements) –, they need to collect both goods in equal quantities. Exchange for one good in one relation thus promotes transfers in other relations providing the complementary good. Each actor will engage in a series of dyadic bargaining sessions to obtain as much as he can from each good. However, this procedure is not suited to define mixed networks in a proper sense.

To eliminate this problem, we introduce a new and more precise classification of network connections which, however, embraces pure negatively and pure positively connected network structures as special cases. It is thus ensured that profit point predictions can be compared with empirical evidence stemming from laboratory experiments on negatively and/or positively connected networks. However, it eliminates, on the one hand, the troublesome correspondence between negative connections and the concern for the distribution of just one homogeneous good. And, on the other hand, it abandons the correspondence between positive connections and the pairwise distribution of perfectly complementary goods. The new classification hinges on three fundamental network characteristics.[8]

Before we can define the new relational types, we thus need to introduce these basic network parameters. In any given bargaining network, we can distinguish between peripheral and non-peripheral positions. Specifically, actor i is said to be at a peripheral network position if he has just one bargaining partner (i.e., $n_i = 1$). Otherwise, actor i is said to be at a non-peripheral position if he has two or more distinct bargaining partners (i.e., $n_i \geq 2$).

In addition, we introduce two node-specific parameters for the precise classification of relations. Specifically, let m_i be the number of actor i's bargaining relations (i.e., m_i is the number of (non-directed or symmetric) arcs associated with node i) and let g_i be the number of exchanges actor i intends to complete (i.e., g_i could denote the

7 Note, however, that positive connections do not necessarily involve distinct goods. Exchanges between, say, a professional athlete and his agent as well as complementary relations between the agent and organizers of athletic events may involve money only. Or, a broker can obtain information in one relation which eventually makes it possible to give advice to another of his partners. However, neither of these alternative considerations on positively connected ties solves our methodological problem.

8 Willer and Skvoretz (1999) also present an alternative categorization of network connections which, as they emphasize, differs from the positive/negative classification. Their distinction embraces five types of connections. It draws on the ranking and values of several node-specific parameters which are, however, at least partly, unique to their theoretical approach.

number of goods i wants to buy or sell in a period), where $m_i \geq g_i \geq 1$.[9] For any actor i in a given network, it thus will either hold $m_i > g_i \geq 1$ or $m_i = g_i \geq 1$. These cases refer to the following situations:

- If $m_i > g_i \geq 1$, i has more bargaining ties than he can conclude exchanges. Actor i therefore has a set of competing bargaining ties with respect to the exchange of one or more substitutable goods. He can thus select his exchange partner(s) from his set of exchange relations.
- If $m_i = g_i \geq 1$, i can conclude just as many exchanges as he has bargaining partners. We say that i has non-competing bargaining relations. Consequently, there are two possibilities: $m_i = g_i = 1$ or $m_i = g_i > 1$. When $m_i = g_i = 1$ is given, i has just one bargaining relation with respect to an exchange of one specific good he intends to complete. On the other hand, if $m_i = g_i \geq 2$ is satisfied, i has different non-competing bargaining relations with respect to the exchange of two or more goods which are either complementary or independent (i.e., neither complementary nor substitutable).[10]

Taking into account that an actor may either be located at a peripheral position or at a non-peripheral position, these above cases provide the basis for the new relational classification of an individual i who is matched with a bargaining partner j:

Rival Orientation: i has a rival orientation in the relation with j if

- i has a non-peripheral position ($n_i \geq 2$) and competing bargaining ties from which he can select exchange relations ($m_i > g_i \geq 1$), or,
- i is at a peripheral position ($n_i = 1$) and his only partner j has competing bargaining relations ($m_i = g_i \geq 1$ and $m_j > g_j \geq 1$).

Non-Rival Orientation: i has a non-rival orientation in the relation with j if

- i has a non-peripheral position ($n_i \geq 2$) and non-competing bargaining ties ($m_i = g_i \geq 2$), or,
- i is at a peripheral position ($n_i = 1$) and his only partner j has non-competing bargaining relations ($m_i = g_i \geq 1$ and $m_j = g_j \geq 2$).

Our postulate reflects that people decide about the type of their connections in reality – we assume that every system actor who has two or more ties in the given negotiation network classifies each of his relations as either a rival or a non-rival connection. This

9 Our postulate reflects that these parameters denote decisions and/or restrictions people take or face in reality. However, in experimental work the parameters m_i and g_i are exogenously fixed by the experimenter, together with the relations defining the bargaining structure.

10 Note that here lies a major difference between our classification and the negative/positive connection classification. A positive connection is generally identified with perfectly complementary goods. Our classification is flexible enough to embrace situations in which the analogous of a positively connected tie does not necessarily ask for complementary goods to be exchanged.

typology further draws on the assumption that an actor with just a single bargaining partner (i.e., an actor at a peripheral position) in a network with two or more relations simply adjusts to the relational categorization of his only partner (in the special case of the Dyad-structure, each actor has a peripheral position and, by postulate, a non-rival orientation).[11] To keep things as simple as possible, however, the actors' assessments of relations are exogenous components in our model.[12]

Combining the possible cases of rival and non-rival orientations for two adjacent actors *i* and *j* in any exogenously given bargaining network suggests a parsimonious classification of relations:

Pure Rival Connection: A pure rival relation between *i* and *j* exists if both actors have a rival orientation.[13]

Pure Non-Rival Connection: A pure non-rival relation between *i* and *j* exists if both actors have a non-rival orientation.[14]

As a consequence of our assumptions, a peripheral actor never will be involved in a mixed relation. When an actor *i* has two or more partners, however, he may face a non-peripheral bargaining partner *j* with another relational orientation. If so, a mixed relation between *i* and *j* will exist.[15]

Mixed Connections: A mixed relation between *i* and *j* exists if the actors differ in terms of their relational orientation.

Mixed Rival Connection: A mixed rival relation is said to exist if *i* has a rival orientation and *j* has a non-rival orientation.

Mixed Non-Rival Connection: A mixed non-rival relation is said to exist if *i* has a non-rival orientation and *j* has a rival orientation.

---

11 Note that this assumption does not exclude that the individual classifications of network connections simply reflect systemwide incentives. Our postulate therefore should not create problems in experimental work as long as test persons (have learned to) systematically react to incentives. Suppose, in accordance with the usual design of network exchange experiments (e.g., Skvoretz and Willer 1991), that subjects' monetary compensation for participation explicitly depends on their bargaining success. We expect that these incentives ensure, at least after several training rounds, relational assessments in the sense of the experimenters.

12 Following the usual practice in the field, we thus predict exchange profits and structures only. This clearly simplifies the model – apart from the actors' structural positions (e.g., just one potential exchange relation or several bargaining relations), the assessments of relations may reflect the network type (e.g., dating or communication network), the number and type of goods to be divided, and the existence of systemwide restrictions (e.g., laws, norms, rules).

13 If the bargaining network is characterized by $m_i \geq g_i = 1$ for all $i = 1, 2, \ldots, n$, the structure of pure rival connections corresponds to what is normally referred to as a pure negatively connected network (with a one-exchange rule).

14 The only experimental work on a positively connected network structure is due to Yamagishi, Gillmore, and Cook (1988). This bargaining network was characterized by $m_i = g_i = 2$ for all non-peripheral actors, and thus constitutes a special case of a pure non-rival network.

15 Under special parameter constellations, these mixed relations correspond to negatively-positively and positively-negatively connected bargaining structures, respectively.

More precisely, there is a mixed rival connection if $n_i \geq 2$ and $m_i > g_i \geq 1$ as well as $n_j \geq 2$ and $m_j = g_j \geq 2$, but a mixed non-rival connection if $n_i \geq 2$ and $m_i = g_i \geq 2$ as well as $n_j \geq 2$ and $m_j > g_j \geq 1$. In all other cases, the relational orientations of two bargaining partners coincide such that they face either a pure rival or a pure non-rival connection.

Consider the example of the basic T-Shape structure where position B is connected to two peripheral actors, A_1 and A_2, as well as to position C. The latter is in addition tied to the peripheral position D. Assume further that we seek to ensure mixed relations by inserting complementary goods at both ends. Assume that the peripheral A's are provided with good *a* and the peripheral position D with good *d*. According to Yamagishi, Gillmore, and Cook (1988), the B–A ties is negatively connected – the A's are competitors in providing B with good *a* – while the B–C and C–D relations are positively connected. The theoretical definition of negative and positive ties, however, creates difficulties in appropriately defining such a mixed network. Our new classification of network connections makes no assumption about whether ties classified on the dyadic level would promote or prevent transfers in other relations. We thus can uniquely define mixed networks. In our example, B $(n_B = 3,\ m_B = 3,$ and $g_B = 2)$ as well as the two As $(n_A = 1, m_A = 1,$ and $g_A = 1)$ have rival orientations, due to $m_B > g_B \geq 2$ and $m_A = g_A \geq 1$. On the other hand, both C $(n_C = 2, m_C = 2,$ and $g_C = 2)$ and D $(n_D = 1,$ $m_D = 1,$ and $g_D = 1)$ have non-rival orientations since $m_C = g_C \geq 2$ and $m_D = g_D \geq 1$. This results in a T–Shape with mixed connections, where the B:A relations are pure rival connections and the C:D is a pure non-rival connection. The B:C relation is, from B's point of view, a mixed rival connection while C would see the connection as a mixed non-rival one. Consequently, both B and C can engage in two exchanges per round while the As as well as D can only exchange once. Since B has three ties but seeks to exchange twice, he can choose between exchanging with A_1 and A_2, or with C and either of the As. Actor C, in contrary, is – if two exchanges need to be concluded – forced to exchange with C and B. Note, however, that the sequence of dyadic exchanges does not matter for our new classification of relations. It therefore allows for a proper definition of a mixed network.

11.3 Theoretical model

Consider an exogenously given network with *m* mutual ties between a finite number of rational actors (*i*, *j*, *k*=1, 2, . . ., *n*).[16] These symmetric relations limit the

16 As will become clear later, the conclusion of the model does not rest on the assumption of fully rational and selfish egoists. As long as an actor's utility is determined by his own share of the cake and the actor purposively negotiates his share of the surplus, it suffices to assume boundedly rational, but learning actors. The use of the rational actors' framework, however, simplifies the presentation of the model.

matches of potential partners for negotiations and exchanges. Each bargaining session refers, by postulate, to the bilateral distribution of a fixed quantity of a perfectly divisible resource (e.g., money). Exchange appears here, in accordance with sociological approaches (e.g., Bonacich and Friedkin 1998; Willer 1999), as an agreement of two rational actors on the division of a fixed surplus. Specifically, we assume that the actors i and j bargain over the partition of a surplus of given value $v_{ij} = v_{ji}$. When x_{ij} represents i's negotiated share of the value v_{ij}, it holds that $0 \leq x_{ij} \leq v_{ij}$.[17] Put differently, x_{ij} denotes i's negotiated exchange profit in the relation with j.

The profit shares x_{ij} and x_{ji} are to be explained in terms of structure and positional features. For that purpose, it is postulated that, once i and j negotiate over the partition of the value v_{ij}, they determine their profit shares as if they would apply the generalized version of the Nash (1950, 1953) bargaining solution from cooperative game theory (see, e.g., Binmore 1987, 1992). That is, they choose the profit shares x_{ij} and x_{ji} as if they would solve the optimization problem

$$\max x_{ij}^{b_i} \; x_{ji}^{b_j} \text{ subject to } x_{ij} + x_{ji} = v_{ij}, \tag{11.1}$$

where the positive parameters b_i and b_j refer to i's and j's absolute level of individual bargaining power.[18] As will become clear below, the solution of the optimization problem implies that the split of the given surplus between i and j reflects the combination of their bargaining powers (i.e., the distribution of "relative bargaining power" determines the negotiated exchange profits). And, most important, the generalized Nash bargaining solution has a solid non-cooperative foundation in the limiting equilibrium of Rubinstein's (1982) Alternating Offers Game.[19] Even though our approach has desirable properties, it differs from other theories on exchange

17 If the actors perpetually disagree, they do not get a proportion of the surplus. That is, the payoff associated with disagreement is 0 for both network partners.

18 Binmore (1992: 184–188) proves that the solution of the optimization problem specified in eq. (11.1) is the only bargaining solution which satisfies the following three axioms: (A) the bargaining outcome does not depend on how the negotiation partners' utility scales are calibrated; (B) the bargaining outcome is individually rational and Pareto- efficient; (C) the actors' choice is independent of the availability or unavailability of irrelevant alternatives (i.e., if the bargaining partners sometimes agree on a specific outcome when another outcome is feasible, then they will never agree on the latter when the former is feasible).

19 The AOG is a non-cooperative game in which rational egoists alternate in making proposals on how to divide a cake with one time periode elapsing between offers. If the focus is on a scenario in which the amount of time between proposals vanishes, its equilibrium coincides with the generalized Nash bargaining solution (cf., Binmore 1985, 1998; Muthoo 1999; Osborne and Rubinstein 1990). This insight is especially important if one agrees with Nash (1950, 1951, 1953) and regards non-cooperative games more fundamental than cooperative ones. It is noteworthy that (if theoretical predictions should be tested against experimental results) the subgame perfect Nash equilibrium of the limiting Rubinstein game is robust against deviations from the strict logic of offers and counteroffers as can happen in computerized experiments (Perry and Reny 1993).

networks. While Bienenstock and Bonacich (1992, 1997) conceptualize a network of potential exchange partners as a cooperative game with transferable utility, they predict negotiation profits via other solutions concepts from cooperative game theory (e.g., core, kernel). Other theorists (e.g., Lovaglia et al. 1995; Skvoretz and Willer 1993) determine the bilateral bargaining outcome by defining and equating "resistance" equations. Although they do not use game-theoretic concepts, resistance theorists (e.g., Willer 1999) emphasize that their approach rests on the premise of strategically rational actors. Still others (e.g., Yamagishi and Cook 1992; Friedkin 1993; Yamaguchi 1996) more or less explicitly refer to non-strategic profit-maximizing behavior when they model exchange relations in networks.

While we think that the foundation in non-cooperative game theory should be a relevant aspect of a theory of network exchange, there are yet other reasons for employing the Nash model over existing theories of sociological exchange theory:

- Bienenstock and Bonacich's (1992, 1997) cooperative solution of the core (or the kernel) neglects the strategic process by which players form coalitions and make demands. From the perspective of non-cooperative game theory, however, it seems necessary to ask for a description of precisely this strategic process. Up to now, the core has no grounding in non-cooperative bargaining theory. From an empirical point of view, on the other hand, it weighs heavy that the core measure does not allow for point predictions for all networks which have so far been experimentally studied.[20]
- Yamaguchi (1996, 2000) models the bargaining and exchange game via the assumption of a perfectly competitive market (i.e., an anonymous and centrally located market with many relatively homogenous participants who offer and buy sufficiently homogenous resources only at the systemwide equilibrium price) in which all the price adjustments are costlessly made by a fictitious and neutral Walrasian auctioneer. His approach is a general equilibrium model (viz., each actor acts as if he would solve a constrained CES-utility maximization problem). Hence, Yamaguchi does not allow for strategic rationality, but focusses on dominant strategies only. And, it may be doubted whether this approach is appropriate for the small social settings usually being studied in network exchange experiments.
- A group of prominent models (e.g., Willer and Skvoretz's (1993) ER-model or Lovaglia et al.'s (1995) GPI-RD model) is based on (extensions of) the resistance logic. The resistance logic of these approaches, however, is just an ad hoc transformation of Heckathorn's (1980) original resistance equation with additional but questionable assumptions about the actors' aspiration levels and

20 For example, its application gives only range predictions for a structure as elementary as the 4-Line network. Moreover, the core measure predicts unstable exchange outcomes, that is, it has an empty core, for, for instance, the simple Kite structure (see Figure 11.3 for both networks). It thus contradicts experimental work (cf., Willer 1999) which demonstrates just the opposite.

conflict points. Moreover, these theories do not offer a measure which captures the effects of structural position on bargaining power.

– Heckathorn's (1980) original resistance model, however, corresponds to the Kalai-Smorodinsky (1975) bargaining solution (the above mentioned models are not sufficiently characterized by the Kalai-Smorodinsky axioms). Nevertheless, the Nash bargaining solution is to favor over the Kalai-Smorodinsky solution for at least two reasons. First, the latter does not possess a realistic non-cooperative foundation. There is a non-cooperative game invented by Moulin (1984) which is often cited as providing a non-cooperative foundation for the Kalai-Smorodinsky solution. This game, however, has nothing to do with either network exchange experiments nor real life bargaining situations. Among other things, Moulin's game requires the presence of an impartial referee who costlessly organizes consecutive lotteries about feasible proposals in which all the players voluntarily engage. Second, the Kalai-Smorodinsky solution is most relevant in a situation in which the surplus to be divided between two players contracts or expands in a way that possibly makes the set of allocations asymmetric. In network exchange experiments, however, the cakes to be partitioned neither shrink nor grow.

The concrete application of the generalized Nash bargaining solution, even though favorite to competing approaches, requires a numerical specification of each actor's bargaining power. The latter is exogenous to the Nash bargaining model. Therefore, it is precisely here where the basic idea of sociological theories for exchange networks comes in – we assume that, once a relation has been classified as either a rival or a non-rival connection, each actor's bargaining power results from his structural position in the network under consideration. Therefore, our model essentially combines the generalized version of the Nash bargaining solution with a specific definition of each actor's bargaining power in terms of relational features and structural embeddedness.

More precisely, we claim that actors, depending on their structural positions in the exogenous bargaining network and the values of their relations (i.e., the sizes of the cakes to be partitioned), differ in terms of their "network control" (i.e., the extent to which an actor controls the relations to him by his relations to others). And, depending on the actor's assessment of the relation he has with a potential exchange partner (i.e., rival or non-rival), his network control either positively or negatively affects his individual bargaining power. Once individual bargaining powers have been determined, the generalized Nash bargaining solution enshrines the distributions of relative bargaining power and exchange profit in the relations under consideration.

After having introduced the new classification of relations and the generalized Nash bargaining solution, it is left to show how a given negotiation structure determines network control and how it is then related to bargaining power.

11.4 Network structure and bargaining power

Consider an exogenously given bargaining network with a fixed number of structural positions and given values of their respective ties (i.e., the size of the cakes to be partitioned). Such a network may reflect a snapshot of a negotiation situation in real world where cake sizes represent surpluses to be negotiated. A positive surplus may materialize if the buyer's willingness to pay exceeds the seller's reservation price. Or, it simply represents a bargaining structure implemented by the experimenter in a laboratory experiment where each tie refers to a fixed amount of points available for negotiation. Starting from such a network, we postulate that actors, depending on their structural position and the size of the cakes they can negotiate on, differ in terms of their network control. And, depending on an actor's relational assessment of ties to adjacent partners (i.e., rival or non-rival) and the number of exchanges he intends to complete, his network control positively or negatively affects his individual bargaining power. We now successively present these assumptions and their implications in detail.

11.4.1 Negotiation structure and network control

11.4.1.1 Unequally valued relations

Let the $n \times n$ matrix $\mathbf{V}$ with main diagonal elements $v_{ii} = 0$ for all i and off-diagonal elements $v_{ij} \geq 0$ for all $i \neq j$ represent the exogenously given network of m valued bargaining relations between the n actors. While the relation between the bargaining partners i and j is always symmetric, an actor's relations with distinct partners may differ with respect to the values at stake – the corresponding off-diagonal elements $v_{ij} = v_{ji}$ express whether i and j are bargaining partners and, if so, how large the cake is they can divide. Formally, it holds $v_{ij} = v_{ji} > 0$ in the presence of a bargaining relation between $i \neq j$, but $v_{ij} = v_{ji} = 0$ in its absence.[21]

Even if matrices of valued adjacencies differ, they may represent the same relational structure. A standardization is thus reasonable. Let $\mathbf{R}$ be the $n \times n$ matrix of standardized actor relations such that $r_{ij} := \frac{v_{ij}}{\sum_{k=1}^{n} v_{kj}} \geq 0$ for all i,j, and $\sum_{k=1}^{n} r_{kj} = 1$ for all j. That is, $\mathbf{R}$ is the column-stochastic matrix derived from the valued graph. Its off-diagonal element r_{ij} measures i's fraction of the systemwide valued relations to j. In other words, r_{ij} represents i's degree of "control" over the valued relations to

21 In the basic scenario, each pair of connected actors bargains over the partition of just one cake with a specific size. In a slightly more complicated case, there may be more than just one reciprocated tie between each pair of connected actors in the network (e.g., each dyad can divide two pies per round of negotiated exchanges). If so, the sum of the relevant surpluses determines the off-diagonal elements of matrix $\mathbf{V}$.

j in the system. Put differently, r_{ij} informs on *i*'s control over the total of *j*'s resources available for negotiation. For example, $r_{ij} = 0.333$ means that *i* "controls" one third of *j*'s resources to be negotiated on. As will become clear below, in a negotiation structure with homogeneous cake sizes, the interpretation of r_{ij} is even more straightforward. It is the reciprocal of the number of *j*'s negotiation partners. For instance, $r_{ij} = 0.333$ means that *j* has three negotiation partners of which *i* is one. Regardless of the cake sizes, it holds $0 \leq r_{ij} \leq 1$, where $r_{ij} = 0$ indicates that *i* has no control over *j* (i.e., absence of a tie between *i* and *j*) and $r_{ij} = 1$ reflects that *i* has complete control over *j* (i.e., *i* is *j*'s only bargaining partner).

However, r_{ij} reveals only part of the available information on the tie between *i* and *j*. While the *i*-th row of the matrix **R** informs about *i*'s control over each other actor in the system, the *i*-th column of **R** informs about the other's control over *i*. That is, r_{ji} denotes *j*'s control over the valued relations to *i*. Adding up the relevant pairwise elements of **R** defines the network control of actor *i:*

$$c_{ij} := \sum_{k=1}^{n} r_{ik} r_{ki} \text{ for all } i. \tag{11.2}$$

Put verbally, c_i is the degree to which *i* controls the valued relations to him by his valued relations to others. For example, $c_i = ¾$ means that actor *i* controls, via his valued relations to others, ¾-th of their valued relations to him. The control fraction c_i thus may be interpreted as *i*'s 'structural autonomy' as well.[22] And, given information about *i*'s valued relations and those of his partners, its calculation is straightforward. As a consequence, we do not have to assume that every actor has complete information about the overall shape of the network structure. For the determination of the control distribution in the system, it suffices to postulate that everyone has complete information about his own valued relations and those of his network partners. In this regard, we need not assume fully rational but boundedly rational actors only.

11.4.1.2 Equally valued relations

The weak informational requirements reflect the parsimony of our operationalization of network control. However, there is only one experimental study which systematically varies the size of the cakes to be partitioned in bilateral negotiations (Bonacich and Friedkin 1998; see section 11.5.3). All other laboratory experiments

22 The definition of network control (cf. eq. (11.2)) shows that the c_i's are the positive main diagonal elements of the $n \times n$ matrix $\mathbf{C} := \mathbf{RR}$. That is, $c_i = c_{ii} > 0$ for all *i*. Like **R**, **C** is a column-stochastic matrix. That is, $0 < c_i \leq 1$ and $1 - c_i = \sum_{k \neq i} c_{ki} \geq 0$ for all *i*. Since the upper bound of c_i is 1, its complement $1 - c_i$ measures *i*'s 'structural dependence' (i.e., the degree to which the other system members affect, via their valued relations to one another, the valued relations to actor *i*).

focus on networks with equally valued relations. That is, each bilateral bargaining session concerns the division of an identical surplus $v_{ij} = v_{ji} = v$ for all $i \neq j$. Given the current lack of evidence on exchange systems with heterogenous cake sizes, it seems useful to demonstrate the simple calculation of c_i for networks with equally valued relations (viz., applications in section 11.5.1 and 11.5.2).

Let $v_{ij} = v_{ji} = v$ for all $i \neq j$. The $n \times n$ matrix **V** thus reduces to the $n \times n$ adjacency matrix **A** with main diagonal elements $a_{ii} = 0$ for all i and off-diagonal elements $a_{ij} \in \{0,1\}$ for all $i \neq j$ representing the exogenously given and symmetric bargaining relations. More precisely, a_{ij} is a binary measure for the absence or presence of a mutual tie between the actors i and j (i.e., $a_{ij} = a_{ji}$ is coded as 0 or 1 for all $i \neq j$).

Let again **R** be the $n \times n$ matrix of standardized actor relations such that $r_{ij} := a_{ij} / \sum_{k=1}^{n} a_{kj} \geq 0$ for all i, j, and $\sum_{k=1}^{n} r_{kj} = 1$ for all j. In the case of equally valued relations, **R** is the column-stochastic matrix derived from the adjacencies bearing the properties as described above.

The calculation of c_i is then straightforward. A closer look at eq. (11.2) shows that c_i may be alternatively expressed as the mean of the i-th row in the matrix **R**. In other words, i's network control is the mean of i's control over the systemwide relations to his partners. Formally,

$$c_i = \frac{1}{n_i} \sum_{k=1}^{n} r_{ik} = \frac{1}{n_i} \sum_{k=1}^{n} \frac{a_{ik}}{m_k} \text{ for all } i. \tag{11.3}$$

where n_i denotes the number of i's negotiation partners and m_i the number of i's ties in the network structure under consideration.[23] Since the number of positive elements in the i-th row in **R** is always n_i, the relevant control distribution can be practically read off from the standardized actor relations.

The number of i's bargaining partners, n_i, coincides with the number of positive elements in the i-th row of the adjacency matrix **A**, too. As indicated by the far right-hand side of eq. (11.3), information about i's network relations and the number of his partners' ties therefore suffices for the computation of i's network control as well. Put differently, the distribution of network control can be determined when each actor knows his connections and the connections of his partners.

The concept of network control thus requires only weak assumptions about the structural information of network members. This becomes more obvious if c_i is

23 In the basic scenario of an adjacency matrix, the number of each actor's negotiation partners n_i is equal to the number of his ties m_i (i.e., $n_i = m_i$ for all i). If, however, more than just one cake per round is to be divided by each pair of connected network members, every actor has more ties than partners (i.e., $m_i > n_i$ for all i). For instance, if i has $n_i = 3$ bargaining partners where two cakes per tie can be divided, we have that $m_i = 6$. If more than one cake per round can be divided between i and j, we assume per definition that i and j are connected by the respective number of ties. In such a case, we can alternatively write $r_{ij} := a_{ij} / \sum_{k=1}^{n} a_{kj} = a_{ij} / m_j \geq 0$ for all i,j.

expressed in still another way. When S_i denotes the set of the n_i bargaining partners of actor i, it is possible to rewrite eq. (11.3) as follows:

$$c_i = \frac{1}{n_i}\sum_{k\in S_i}\frac{1}{n_k} = \frac{1}{(n_i/\sum_{k\in S_i}(1/n_k))} \text{ for all } i. \tag{11.4}$$

Stated differently, i's network control c_i reflects how many negotiation partners actor i and his partners have. Hence, information about the number of i's bargaining partners and the numbers of their partners allows the calculation of i's network control.[24]

A closer inspection of the far right-hand side of eq. (11.4) shows, moreover, that the concept of network control is compatible with a rational actor perspective – c_i is simply the reciprocal of the estimate a rational actor i will have for the mean number of partners of his partners in a given network.[25] It thus holds that, for example, $c_i = \frac{3}{4}$ expresses that the average number of bargaining partners of i's partners is $\frac{4}{3} = 1.333$. Clearly, i's network control c_i decreases if the mean number of partners of i's partners increases. And, c_i rises if the mean number of bargaining partners of i's partners falls. In the limit case, the mean number of bargaining partners of i's partners equals 1 (i.e., actor i is their only negotiation partner) such that i has full network control $c_i = 1$. So, when the assumption is made that i's behavior reflects his network control or structural autonomy, it is postulated, in effect, that he takes account of the mean number of partners of his partners.

Whether or not we focus on bargaining structures with equally valued relations, the weak informational requirements reflect that i's network control captures the structure only two steps from i (but neglects structural effects which are three or more steps away). The degrees of network control thus may be seen as parsimonious indicators for the actors' structural positions. They are, by postulate, essential determinants of the individual bargaining powers. That is, actor i's network control c_i affects b_i, i's level of individual bargaining power. The direction of the relationship between network control c_i and bargaining power b_i depends, however, on i's categorization of the respective network relation.

Referring to the typology of relational orientations introduced in section 11.2, we now can specify how network control affects individual bargaining power for

24 For the determination of the distribution of network control, we thus do not have to assume that every actor has complete information about the overall shape of the network structure. It suffices to postulate that either everyone has complete information about his own relations and his partners' ties or that everyone knows the number of his bargaining partners and the number of the partners' partners.

25 Feld (1991) explains why friends always seem to have more friends than oneself. For the scenario in which this 'class size paradox' is fully understood, he derives the appropriate estimate for the mean number of friends of an individual's friends. The latter corresponds with the denominator of the far right-hand side of eq. (11.4).

given specific relational orientations. For that purpose, imagine first a system with non-rival relations. Such a network has two essential features: depending on the values involved, *i*'s bargaining relations are more or less substitutable and compete with the relations of *i*'s partners to others ("friends of friends are enemies").[26] Both features suggest that *i*'s absolute bargaining power rises with his network control – by definition, more control means that *i* depends less on his current negotiation partners for exchange and that *i*'s bargaining partners tend to have fewer and/or less valued relations. For the case of rival relations, it thus can be assumed that *i*'s bargaining power b_i rises with *i*'s network control c_i.

It is reasonable to postulate just the contrary for the opposite scenario of non-rival relations. The features of non-rival relations thus suggest that *i*'s bargaining power increases if *i*'s network control decreases. If an actor's structural autonomy is lower, the others more affect, via their valued links to one another, the valued relations to him. Put differently, if his network control is smaller, an individual has, by definition, a higher structural dependence. Especially in the (experimentally relevant) case where different goods flow through the network, this creates additional opportunities in a non-rival setting. Due to the others' transactions, a less autonomous actor may serve as a broker – that is, an actor who either crucially controls the flow of a resource through the system or who controls a specific resource in the system (e.g., a subject who receives one resource for input into the system).[27] In such a setting, relations with different partners are complementary for the resource flow through the system, and concluded exchanges promote transfers with others. The exchange partners of an individual's partners thus appear as intermediaries ("friends of friends are friends"). For the case of non-rival relations, it thus can be assumed that *i*'s bargaining power b_i rises if *i*'s network control c_i falls.

11.4.2 Relational assessments and bargaining power

Drawing on the typology of relational orientations introduced in section 11.2 and taking into account that actors may differ in terms of the number of exchanges they

26 In extremum, this refers to a situation in which bargaining sessions with different partners concern the distribution of just one homogeneous good (what is elsewhere termed a negatively connected network). As a matter of fact, if there are at least two distinct resources, negative connections are also possible – negotiation partners then have to be willing to substitute one good for the other on a one-to-one basis (perfect substitutes).

27 Non-rival relations may be characterized by the property that exchanges with distinct partners increase individual benefits. Such a situation exists, for example, if the focus is on pairwise distributions of different goods which are always consumed together in fixed proportions (perfect complements) – since every actor only cares about the combination of those goods, he will successively engage in a series of dyadic bargaining sessions to obtain as much as he can from each good. Note, however, that non-rival relations do not necessarily involve distinct goods.

wish to complete, we now show how an actor's bargaining power comes about. To capture the effect of potential variations in the number of desired exchanges across the network, we model two plausible hypotheses: actor i's bargaining power b_i increases (decreases) if, ceteris paribus, i has a rival (non-rival) orientation and wants to complete fewer (more) exchange relations. Put differently, the number of exchanges a position needs to conclude can weaken a rather beneficial network position but can also improve the bargaining power of a rather weak structural position.

In detail, we start from the postulate that the number of desired exchanges (g_i) matters at certain positions more than at others. More precisely, we introduce a case distinction in the following definition of a node-specific weight z_i:

$$z_i := \begin{cases} g_i \text{ if } n_i = 1, \text{or}, g_i = 1, \text{or}, m_i = m, \text{or}, g_k = g \text{ for all } k, \\ \text{or}, m_j = m_i, g_j = g_i, c_j = c_i \text{ for at least one } j \in S_i, \\ f_i(g_i) \text{ otherwise} \end{cases} \tag{11.5}$$

where the function $f_i(g_i)$ denotes a Box-Cox (1964) transformation of g_i:

$$f_i(g_i) = \frac{g_i^{d_i} - 1}{d_i} \text{ with } d_i = 1 - 2\frac{m - m_i}{n(n-1)}. \tag{11.6}$$

Put verbally, the node-specific weight $z_i > 0$ either equals g_i or results from a flexible Box-Cox transformation of g_i in which, by postulate, the exponent and denominator d_i falls as $2(m - m_i)/n(n - 1)$ rises. Notice that z_i always increases in g_i, the number of exchanges actor i wants to complete.[28] Since $2(m - m_i)/n(n - 1)$ measures the difference between the network density and i's density contribution (i.e., the density of the bargaining network without i's relations), the Box-Cox scenario assumes that the effect of g_i on the node-specific weight z_i also depends on the number of i's bargaining relations (m_i), the number of system actors (n), and the number of network relations (m). The combination of these additional parameters reduces the positive effect g_i has on the node-specific weight z_i.[29]

Basically, we postulate that actor i's bargaining power b_i increases with his network control c_i when he sees a potential exchange relation as rival. If he classifies a relation as a non-rival one, however, his bargaining power b_i decreases with his network control c_i. To formalize these ideas, we follow Binmore (1985: 273) who defines individual bargaining power as the negative reciprocal of a logarithmic expression.

28 In the Box-Cox scenario, it holds $m > m_i > 0$ by definition. This ensures $0 < d_i < 1$. Therefore, f_i is concave in g_i (i.e., $d f_i/dg_i > 0$ and $d^2 f_i/dg_i^2 < 0$).

29 Since $0 < d_i < 1$ holds in the Box-Cox scenario, f_i is always smaller than the exogenously given parameter g_i.

Taking into account the node-specific weight z_i (and its hypothesized effect on b_i), it is assumed that

$$b_i := \begin{cases} -1/(z_i \ln(wc_i)), & \text{if } i \text{ has a rival orientation} \\ -z_i/\ln(1-wc_i), & \text{if } i \text{ has a non-rival orientation} \\ -z_i/\ln(wc_i), & \text{if } i \text{ has a mixed rival orientation} \\ -1/(z_i \ln(1-wc_i)), & \text{if } i \text{ has a mixed non-rival orientation} \end{cases} . \quad (11.7)$$

where we use the shorthand

$$w := \frac{m+n}{1+m+n} \quad (11.8)$$

The latter is a network-specific fraction which rises with the number of mutual ties in the network, m, and the number of network members, n. In eq. (11.7), the weight w scales the degrees of network control or structural autonomy such that b_i is always a positive number.[30]

Equation (11.5) reflects the general character of our approach. All possible network parameters are taken into account in defining an actor's bargaining power. Of course, the additional parameters play no role when the Box-Cox transformation does not apply. By postulate, it holds $z_i = g_i$ if at least one of the following conditions is met:

- actor i has a peripheral network position (i.e., $n_i = 1$);
- actor i wants to complete just one exchange (i.e., $g_i = 1$);
- actor i is involved in all relations in a network (i.e., $m_i = m$), which, in effect, means that i has the central position in a branch structure (i.e., hub-and-spoke network) of arbitrary size;
- actor i belongs to a bargaining network in which each member wants to complete the same number of exchanges (i.e., $g_k = g$ for all k);
- actor i faces at least one bargaining partner with the same network control, the same number of bargaining relations, and the same number of desired exchanges (i.e., $c_j = c_i$, $m_j = m_i$, and $g_j = g_i$ for at least one $j \in S_i$).

30 An arbitrary choice of w should not be considered since exchange results are of course not robust against choices of the value of w. Starting from this insight, there are three reasons for eq. (11.8), the specific definition of w. First, admissible transformations of network control are those which just change the unit of scale. The weight w clearly fulfills this requirement. Second, any bargaining network may be characterized by the number of mutual ties, m, and the number of system actors, n. It thus is reasonable to define the scaling factor w in terms of these system parameters. Third, weighting should preserve the essential role of network control in our approach. The weight w is a systemwide constant which, at most, moderately changes the original values of network control – because of $m \geq 1$ and $n \geq 2$, it holds that $0.75 \leq w < 1$.

These conditions for $z_i = g_i$ reflect that the application of the Box-Cox transformation does not always make sense.[31] Their inspection shows that $z_i = g_i$ is assumed in either limit cases or situations with some form of symmetry. The limit cases include the scenarios in which i intends to complete just one exchange ($g_i = 1$) or resides at an extreme network position ($n_i = 1$ or $m_i = m$). The symmetric situations embrace the (experimentally relevant) scenario in which the number of exchanges to be completed does not vary across the network (i.e., $g_k = g$ for all k) as well as the case in which i has at least one fully equivalent bargaining partner ($g_i = g_j$ and $m_i = m_j$ and $c_j = c_i$ for at least one $j \in S_i$). For all other situations, however, we assume that $z_i = f_i(g_i)$ such that the weight z_i increases with the number of desired exchanges g_i, but always falls short of it.

Starting from the distribution of network control (cf., eq. (11.2)), each actor's bargaining power results from combining eqs. (11.5–11.8) for either relational classification he may have. We then can derive the negotiation outcomes associated with the different types of relations.[32]

11.4.3 Relational assessments and negotiation outcomes

If i and j negotiate over the partition of the given cake of size $v_{ij} = v_{ji}$, we assume that they determine their profit shares as if they would solve the optimization problem expressed in eq. (11.1). The maximization of the equivalent welfare function $x_{ij}^{b_i}\left(v_{ij} - x_{ij}\right)^{b_j}$ implies that i can obtain the exchange profit

$$x_{ij} = \left(\frac{b_i}{b_i + b_j}\right) v_{ij} = p_{ij} v_{ij} \text{ for } i \neq j \tag{11.9}$$

31 For example, if $g_i = 1$, the Box-Cox transformation would not assign a positive value to z_i. Or, if the number of exchanges to be completed does not vary across the network (i.e., $g_k = g$ for all k), the Box-Cox transformation would not necessarily ensure a common weight $z_i = z$.

32 Empirical evidence on exchange networks, by and large, results from experiments in which a systemwide one-exchange rule holds. Subjects can thus conclude only one exchange per round (i.e., $z_i = g_i = g = 1$ for all i). In such an extreme scenario, eq. (11.7) can be simplified. Substituting $z_i = 1$ for all cases defining b_i in eq. (11.7) facilitates the computation of an actor's bargaining power, such that

$$b_i := \begin{cases} -1/\ln(wc_i), & \text{if } i \text{ has a rival orientation.} \\ -1/\ln(1 - wc_i), & \text{if } i \text{ has a non-rival or mixed non-rival orientation.} \end{cases}$$

Since such negotiation networks show no variation in the allowed number of exchanges, it does not surprise that only network control matters. Consequently, the definition of bargaining power for actors in pure rival (pure non-rival) and mixed rival (mixed non-rival) relations coincides.

where $p_{ij} := b_i/(b_i + b_j)$ defines i's relative bargaining power in the relation with j. And, since $p_{ji} = 1 - p_{ij}$ holds by definition, i's partner j will receive $x_{ji} = (1 - p_{ij})\, v_{ij} = p_{ji}v_{ji}$. Accordingly, the optimal partition of the given surplus depends critically on the combination of b_i and b_j. Put differently, the bargaining power of just one partner is irrelevant for the negotiation outcome – it is the relative bargaining power (i.e., p_{ij} or $p_{ji} = 1 - p_{ij}$) which matters for the profit split.[33] Notice, however, that p_{ij} does not coincide with x_{ij} when $v_{ij} \neq 1$. Since $x_{ij} = p_{ij}\, v_{ij}$ and $x_{ji} = (1 - p_{ij})v_{ij}$, a comparison of actor i's profit share with that of his bargaining partner j yields the following chain of equivalent conclusions:

$$x_{ij} \gtrless x_{ji} \Leftrightarrow x_{ij} \gtrless \frac{1}{2} v_{ij} \Leftrightarrow p_{ij} \gtrless \frac{1}{2} \Leftrightarrow b_i \gtrless b_j \text{ for } i \neq j \tag{11.10}$$

Put verbally, a symmetric distribution of bargaining powers ($b_i = b_j$ or $p_{ij} = 1/2 = p_{ji}$) always yields an equal split of the pie ($x_{ij} = v_{ij}/2 = x_{ji}$). There will be an unequal profit division, however, when the power of the two negotiation partners differs. Specifically, i's exchange profit x_{ij} dominates j's exchange profit x_{ji} such that i gets more than half of the pie if and only if p_{ij} exceeds p_{ji}. Because of $p_{ij} + p_{ji} = 1$, the latter is satisfied if and only if i's relative bargaining power in the relation with j exceeds 1/2. And, this is equivalent to the condition that i's absolute bargaining power b_i exceeds j's absolute bargaining power b_j.

The actors' bargaining powers depend, by postulate, on their structural embeddedness and relational classification. As a consequence, the model implications for the distributions of relative bargaining power and surplus in any given match reflects these determinants. Substituting eq. (11.7) into eq. (11.9), we can distinguish four relational types each of which allows specific conclusions about the effects structure has on relative bargaining powers and negotiated exchange profits.

Pure Rival Connection: If the relation between actors i and j is, from their perspective, a pure rival one, then actor i's profit in match with j is

$$x_{ij} = p_{ij} v_{ij} \left(\frac{z_j \ln(wc_j)}{z_i \ln(wc_i) + z_j \ln(wc_j)} \right) v_{ij} \text{ for } i \neq j. \tag{11.11}$$

Hence, i's relative bargaining power and exchange profit in a pure rival connected relation with j rise, everything else being constant, when either i's network control c_i increases or j's network control c_j decreases. Equivalently, i's relative bargaining

33 As put forward in fn. 24, an actor need not possess complete information about the network as a whole. To calculate c_i it is sufficient that he knows his immediate vicinity, that is, the network structure two steps away from him. However, i's relative bargaining power in match with j, p_{ij}, is affected by c_i as well as c_j. This slightly extends the borders of i's immediate vicinity – to compute his profit share in match with j, i actually requires the knowledge of the network structure three steps away from him.

power and exchange profit increase, everything else again being constant,[34] when either i's number of desired exchanges g_i decreases or j's number of desired exchanges g_j increases.[35]

Mixed Rival Connection: If i classifies the relation to j as mixed rival and j categorizes the relation as mixed non-rival, then actor i's profit in the match with j is

$$x_{ij} = p_{ij}v_{ij} = \left(\frac{z_i z_j \ln(1 - wc_j)}{\ln(wc_i) + z_i z_j \ln(1 - wc_j)}\right) v_{ij} \text{ for } i \neq j. \tag{11.12}$$

For such a mixed rival relational orientation, actor i's relative bargaining power and negotiated profit increase if, everything else being constant, either c_i or c_j rises. Note that conclusions about a reaction of p_{ij} with respect to a change in g_i or g_j are only meaningful if the classification of relations always remains unchanged. However, this is not necessarily guaranteed in case of mixed rival, mixed non-rival, and pure non-rival connections.

Mixed Non-Rival Connection: If i classifies the relation to j as mixed non-rival and j categorizes their relation as mixed rival, then actor i's profit in the match with j is

$$x_{ij} = p_{ij}v_{ij} = \left(\frac{\ln(wc_j)}{z_i z_j \ln(1 - wc_i) + \ln(wc_j)}\right) v_{ij} \text{ for } i \neq j. \tag{11.13}$$

In this mixed non-rival relation i's relative bargaining power and profit share increase, everything else being constant, when either c_i or c_j falls.

Pure Non-Rival Connection: If the relation between actors i and j is, from their perspective, a pure non-rival one, then actor i's profit in the match with j results from

$$x_{ij} = p_{ij}v_{ij} = \left(\frac{z_i \ln(1 - wc_j)}{z_j \ln(1 - wc_i) + z_i \ln(1 - wc_j)}\right) v_{ij} \text{ for } i \neq j. \tag{11.14}$$

Consequently, if their relation is a pure non-rival one, actor i's relative bargaining power and negotiated profit increase, everything else being constant, if either i's network control c_i falls or j's network control c_j rises.

Since $w := (m+n)/(1+m+n)$ and $c_i := \sum_k r_{ik} r_{ki}$ hold by definition while r_{ij} measures i's fraction of the systemwide valued relations to j, and g_i denotes i's desired

---

34 These conclusions reflect that $\partial p_{ij}/\partial c_i > 0$, $\partial p_{ij}/\partial c_j < 0$ and $\partial x_{ij}/\partial c_i > 0$ as well as $\partial x_{ij}/\partial c_j < 0$. The signs of these partial derivatives inform about the reaction of i's relative bargaining power p_{ij} and i's profit share x_{ij} in the match with j when exogenous structural changes affect either i's network control c_i or j's network control c_j, but preserve the valued relation between i and j.

35 These conclusions reflect that $dp_{ij}/dg_i < 0$, $dp_{ij}/dg_j > 0$ and $dx_{ij}/dg_i < 0$ as well as $dx_{ij}/dg_j > 0$. Note, however, that they are only meaningful as long as $m_i - g_i \geq 2$ as well as $m_j - g_j \geq 2$ since otherwise the classification of relations changes if g_i or g_j change.

number of exchanges, we thus may uniquely predict the distributions of relative bargaining power and negotiated profit for any given combination of relational orientations in each exogenously given bargaining network with m valued ties between n actors. A comparison of the four model conclusions shows, moreover, that the negotiated profits associated with the distinct relational types just differ in terms of p_{ij}, i's relative bargaining power in the match with j. And, the calculation of relative bargaining powers and negotiated profits requires, in principle, just a pocket calculator.[36]

Analyses of concrete network structures illustrate, as will become clear below, the straightforward application of our approach. To compare such theoretical predictions with empirical observations, we first need to describe and select relevant experimental studies.

11.5 Applications

There are a bulk of experimental results with regard to profit distributions in simple exchange networks (e.g., Bienenstock and Bonacich 1993; Lovaglia et al. 1995; Skvoretz and Fararo 1992; Skvoretz and Willer 1993; Yamagishi, Gillmore, and Cook 1988). However, all these experiments were subject to a relatively homogeneous experimental protocol. Subjects at distinct structural positions did not differ with respect to the number of exchanges they could complete per period. And, there was no variation in the value of the relationships across the network (i.e., equal cake sizes to be divided in all relations). While there are various theoretical approaches (Bienenstock and Bonacich 1992; Burke's 1997; Friedkin's 1986, 1992; Lovaglia et al. 1995; Skvoretz and Willer's 1993; Yamagishi and Cook's 1992; Yamaguchi's 1996) which make acceptable predictions for a range of such simple networks (e.g., 4-Line, Stem, Kite, or 3-Branch), there is no general theory of exchange networks which is parsimonious and could coherently be apply to complex network structures. While the

36 Note that dyadically negotiated profits need not always be realized since pure rival networks may reflect exogenous restrictions (e.g., one-exchange rule). This may have further consequences for the network structure as a whole. In such settings (and in mixed exchange networks as well), at least one system member may select his actual exchange partners from a larger set of potential exchange partners. Depending on the negotiation outcomes, he may never complete transfers with one or more of his negotiation partners (consistent or permanent exclusion). Braun and Gautschi (2007) provide an analysis of such "network breaks" and give necessary and sufficient conditions for coincidences or deviations between bargaining and exchange structures. Their conclusions on "network breaks" (e.g., in the T-SHAPE, the H-SHAPE, or the X4-LINE structure) are strongly supported by empirical evidence (viz., Markovsky, Willer, and Patton (1988); Simpson and Willer (1999)). Braun and Gautschi (2007) further provide testable conclusions on a wide range of network structures not yet tested in laboratory experiments.

Network Control Bargaining (NCB) approach presented in this paper can easily be applied to such simple networks, this paper demonstrates the generality of our model by submitting its theoretical predictions to three different sets of experiments on complex networks.

Due to theoretical considerations and/or experimental evidence on complex networks, some older models have been successively adjusted or revised to now be able to at least address some aspects of complex networks. We compare NCB predictions for dyadic exchange outcomes in complex networks with such alternative approaches. Since none of the alternative approaches makes predictions for both complex characteristics – a variation in the number of exchanges positions can complete per period, and, variations in the value of the relationships across the network –, we refer to different theories for comparison with NCB predictions on these two characteristics.

With regard to results on networks where certain positions were allowed to complete more than one exchange per round, we compare our results to 'updated' versions of Skvoretz and Willer's (1993) GPI-measure. Those are Markovsky et al.'s (1993) GPI-R with an adapted version of the exchange resistance equation, Willer and Skvoretz' (1999) GPI-I-Resistance measure[37] and Lovaglia et al.'s (1995) GPI-RD measure, however, with a revised interpretation of the degree index.[38] Empirical validation is done using results from Skvoretz and Willer (1993) as well as from Willer and Skvoretz (1999).

For networks with heterogeneous cake size, we use results from Bonacich and Friedkin (1998).[39] NCB predictions are compared to Bienenstock and Bonacich's

37 The updated version of the GPI-R measure re-defines (and introduces a case distinction for) the resistance equation. The precise specification of the resistance equation depends on Willer and Skvoretz' (1999) categorization of network connections which draws on the ranking and values of several node-specific parameters. As a consequence, the resulting GPI-R measure is unique to their theoretical approach on network connections (i.e., inclusive, exclusive, null, inclusive-exclusive, and inclusive-null connections; cf., Willer and Skvoretz 1999: 197–199). Moreover, to account for inclusive-exclusive and inclusive-null connections, an additional adjustment in the calculation of the original Graph-theoretic power index (GPI, Markovsky, Willer, and Patton 1988) is necessary. This results in the new GPI-I-Resistance measure.

38 Calculation of the degree index in the original approach is based on the number of ties negotiating partners possess (i.e., their respective number of alternative negotiation partners). However, if a position is allowed (more than) two exchanges per period, the number of negotiation partners is not necessarily unique anymore. A position's effective number of ties need then be calculated as a weighted average of the number of exchange possibilities a position has for each of its allowed exchanges. Consequently, its effective number of ties depends on the sequence of exchanges the position engages in (for an example, see Lovaglia et al. 1995: 142–143).

39 Unfortunately, Bonacich and Friedkin (1998) do not report these results as profit splits between positions, as is normally the case in the literature. Instead, exchange results are graphically depicted. Since the data are not publicly available, we try to read profit splits off these graphics as good as possible.

(1992, 1997) core measure. This cooperative solution from game theory can be applied rather straightforward to complex networks with differences in the value of the ties if one assumes that the payoff $v(S)$ denotes the maximum value of all cakes the set of actors S in the core can divide. However, there are networks with an empty core. In general, Bonacich and Friedkin suggest to make predictions based on an extension of the Power-Dependence theory (Yamagishi and Cook 1992) which is, under certain assumptions introduced by Bonacich and Friedkin (1998: 162–164), identical to the kernel. Since Friedkin's (1993, 1995) Expected-Value theory can also deal with heterogeneous cake sizes, we especially compare NCB predictions to those of these two models. For a concise description of the above mentioned models for exchange networks, we refer the reader to the original literature. Before we move on and confront our theoretical predictions with empirical evidence from complex networks, it is advisable to say a few words about network exchange experiments in general. The design of experiments on exchange networks has common features (see, e.g., Skvoretz and Willer 1991). All experiments consist of several rounds of negotiation and exchange, while the endogenously given relational structure is kept constant. Bargaining sessions involve adjacent network positions only, where usually a cake of identical size (normally 24 "profit points") is to be split in any bilateral match. Experiments generally concern negatively connected networks with a one-exchange rule. That is, the number of exchanges per connection and round is restricted to one. While the latter two points are relaxed in the experiments to be discussed, the experimental protocol otherwise remains unchanged. That also includes that negotiations between adjacent positions occur as a series of offers and counteroffers. Negotiations stop when an agreement is reached.[40] Bargaining experiments on exchange networks thus reflect the non-cooperative scenario of Rubinstein's (1982) Alternating Offers Game (respectively its limiting solution, see fn. 19). This game has, as this paper made clear, a cooperative solution in the presented generalized Nash bargaining solution (Nash 1950, 1953).[41]

40 Partly due to a computerized setting, proposals can be made within seconds and bargaining sessions do not last long (viz., agreement in less than a few minutes).

41 However, the theoretical assumption of rational and selfish actors need not hold in experimental situations. Student populations need time to learn to rationally play the game, that is, to systematically exploit their strategic position in the network and to maximize exchange profits. Fortunately, Young (1993, 2001: ch. 8) shows that the full rationality assumption is not needed to deduce the generalized Nash bargaining solution. He shows that the generalized Nash bargaining solution is "stochastically stable" (i.e., robustness under small, persistent random shocks (Foster and Young 1990)). That is, the high-rational solution from game theory has a representation in a low-rational environment through the process of learning. Put differently, whether we have hyper-rational actors who jump into equilibrium right away or whether boundedly rational actors (who are allowed to make minor mistakes from which they learn) adapt to their environment and learn from past outcomes, does not matter. In equilibrium, both sorts of actors will split the cake as if they would implement the generalized Nash bargaining solution.

The experimental protocol knows no misdirection. Experimental subjects (normally undergraduates who participate for pay) receive general information about the purpose and the number of rounds of the experiment. They also possess complete information about the bargaining rules, the earnings of their partners, the shape of the negotiation structure, and their own positions within the network. Therefore, experimental results are interpreted as effects the given network structure has on exchange patterns and/or profit divisions between adjacent positions. And, results on exchange profits are represented by the means of profit points the advantaged positions in given matches could realize over several rounds of the experiment. Mostly, this is the mean over the last 10 rounds of the experiment while the first 10 rounds are discharged as learning and adaption rounds.

11.5.1 Branch-NMQ structures

Starting point of an exchange network experiment is a specific bargaining structure which limits matches between potential exchange partners. The presentation of experimental findings therefore refers to the types of structural positions (A,B,C,D,E) the different actors have. Individuals located at structurally equivalent positions are normally distinguished by numeric subscripts (e.g., A_1, A_2). In this subsection, we exclusively focus on branch structures where a central actor is connected to a specific number of peripheral actors (hub and spoke network). We denote the branch structure under investigation, in accordance with Willer and Skvoretz (1999), as Branch-*NMQ* structures. *NMQ* refers to node-specific parameters which are, at least partly, unique to their theoretical approach and which are the basis for an alternative categorization of network connections (p. 197–200). *N* is the number of exchange relations connected to a node (i.e., the number of potential exchange partners), *M* is the maximum number of relations in which the node can benefit, and *Q* denotes the minimum number of relations in which exchanges must be completed.[42] Therefore, *Q* is a subset of *M* while the latter itself is a subset of *N*.

N and M are thus equivalent to n_i and g_i, respectively, in our theory. Therefore, (N – M) gives the number of exchange partners which, in every round of bargaining and exchange, are necessarily excluded from exchange. However, since the parameter Q is unique to Willer and Skvoretz' (1999) theory, we cannot compare empirical results for networks with $M \neq Q$ to NCB predictions. Take a look at Figure 11.1. NCB makes predictions for Branch311 and Branch333. However, predictions for Branch332 cannot be made by NCB since $M > Q$ (a dotted circle indicates $M = 3 > 2 = Q$). Only the

42 In network exchange experiments, subjects with $Q < M$ are forced to exchange at least Q times and at most M times. When fewer than Q exchanges with adjacent partners are completed, no points at all, not even from completed exchanges, are paid out.

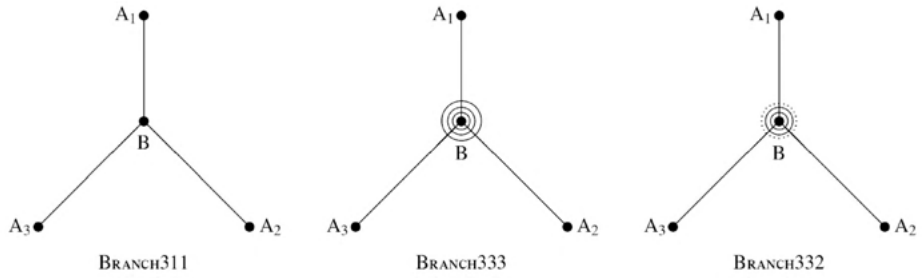


Figure 11.1: Example of complex Branch-NMQ structures according to Willer and Skvoretz (1999)

experiment will tell whether the central actor B will eventually conclude two or three exchanges. It thus cannot be decided in advance whether the negotiation structure Branch332 in fact coincidences with a Branch322 or a Branch333 exchange structure. To test NCB predictions against experimental data from Branch-NMQ structures and predictions from GPI-R (Markovsky et al. 1993) with an adapted version of Exchange Resistance, where the definitions of the ER equations depend on the combination of N, M, and Q (see Willer and Skvoretz 1999: 2005), we restrict ourselves to Branch structures where M = Q hold. Table 11.1 reports the results.

It can be seen that NCB makes rather accurate predictions for the profit points of B in relation with A in all eight branch structures. Note that the first four branch structures in Table 11.1 are pure non-rival ones while the latter four branch structures are pure rival ones. NCB predictions are less off the empirical observations than predictions from GPI-R in all relations. This is also reflected by the two goodness-of-fit measures we report in Table 11.1.

The branch structure is a network where a central monopolist B can exploit his peripheral exchange partners since he can play the As off against each other (put literally, he can, up to a certain point, threaten the As with permanent exclusion from exchange). It can thus be expected that B's profit increases, ceteris paribus, in the number of exchange partners but decreases, ceteris paribus, in the number of exchanges he must complete. NCB and GPI-R predictions as well as the experimental results clearly show this latter fact.[43] For instance, the more exchanges B must conclude, the lower his profits, as the comparison of Branch533 to Branch555, or, Branch311 to Branch322 clearly shows. The most dramatic effect, that is, the complete loss of the monopoly rent can be found if B must conclude exchanges with all of the A's (i.e., N = M or $n_i = g_i$). In Branch333, B only gets about one third of the profit while in Branch332 he could still get about five sixth of the profit. The same

43 For evidence of the former claim, see Braun and Gautschi (2006).

Table 11.1: Observed and predicted dyadic profit splits in complex branch structures.

Network	Match	Tie[a]	NCB	GPI-R	Observed (s.e.)
Branch333	B:A	pnr	7.97†	8.92†	7.96 (0.61)
Branch444	B:A	pnr	7.37†	8.19†	7.53 (0.50)
Branch555	B:A	pnr	6.95	7.62	5.70 (0.34)
Branch777	B:A	pnr	6.39	6.82	5.09 (0.23)
Branch311	B:A	pr	21.65†	23.00	21.63 (0.49)
Branch322	B:A	pr	19.73†	no prediction	19.62 (0.36)
Branch533	B:A	pr	20.80	no prediction	19.71 (0.30)
Branch755	B:A	pr	20.68	no prediction	16.45 (0.25)
AD[b]			0.55 (1.02)	1.33	
MD[b]			0.36 (0.95)	0.63	

Note: Observed profit splits as reported in Skvoretz and Willer (1999). NCB = Network Control Bargaining model. Predictions and observations are for profit points (out of a cake of 24 profit points) of the structural position B. GPI-R (Markovsky et al. 1993) with an adapted version of Exchange Resistance, where the definition of the ER equations depends on the combination of *N*, *M*, and *Q* (no predictions for inclusive-exclusive and inclusive-null), see Willer and Skvoretz (1999).
[a]Classification of relations: pnr = pure non-rival; pr = pure rival.
[b]AD = Absolute Deviation (the sum of absolute distances between observed and predicted profit points relative to number of comparisons); MD = Mean Deviation (the Euclidean distance between observed and predicted profit points relative to number of comparisons). Figures in brackets for NCB refer to the comparison of ties with the latter three networks excluded.
†Daggers indicate that predicted values fall within the 95% confidence interval of the observed values. Put differently, these predictions fit the observations at the $p < 0.05$ significance level (two-tailed tests).

dramatic collapse in profit points for B can be observed for the remaining (N = M)-branches in Table 11.1, that is, for the Branch777 in comparison to the Branch755.

NCB seems to be able to make rather accurate predictions for a first set of complex networks. However, branches are rather simple structures. The following two subsections will use network structures with a little more complexity as the measure rod against to pitch our theory.

11.5.2 Diamond, H- and T-Shape structures

The negotiation networks to be discussed in this subsection basically exhibit the same property as the branch structures just discussed: certain positions must exchange more than once. Figure 11.2 depicts the negotiation networks. For instance, in the H-2-3 network, position C must conclude two exchanges. He can, however, choose from his exchange partners A_1, A_2, and D. On the other hand, position D must exchange three times. Since he only has three possible negotiation partners, he is forced to conclude one exchange per round with each adjacent position. D

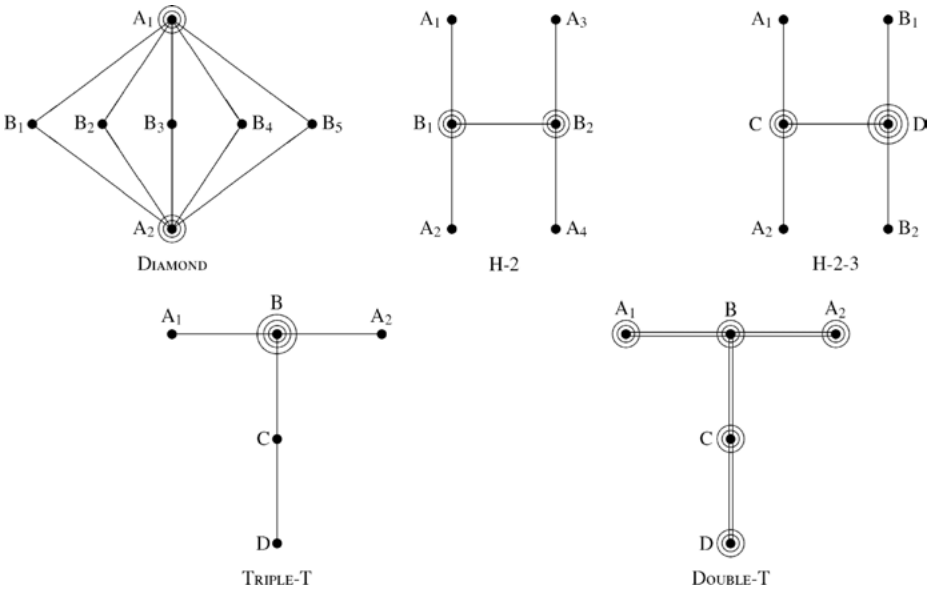


Figure 11.2: Examples of complex structures according to Willer and Skvoretz (1999; Diamond, H-2-3) and Skvoretz and Willer (1993, H-2, Triple-T, and Double-T).

thus is in a rather weak negotiation position and it is to be expected that he can be exploited by his partners. Generally, the experimental results should thus show that the advantage of a structural position in a pure or mixed rival relation is weakened by an increasing number of necessary exchanges.

In addition to the number of exchanges per negotiation round, Figure 11.2 introduces yet another feature. A position who must exchange more than once can, but need not, engage in multiple exchanges with an adjacent position. In the Double-T network, the number of lines connecting two adjacent positions indicates the maximum possible exchanges per round between these positions. Position B, who must conclude two exchanges per round, can thus choose to exchange with just one adjacent position (A_1 only or A_2 only) or with two distinct positions (viz., A_1 and A_2, C and A_1, or C and A_2). While the number of exchanges to be concluded, ceteris paribus, decreases the profit points of the focal positions (the Double-T structure is a pure rival one), it can be expected that the number of cakes which the focal position can divided with one partner, ceteris paribus, increases his bargaining power (and thus his profit points). A position can then threaten more of his partners with permanent exclusion from exchange – due to the possibility of multiple exchanges in one relation – than in a similar situation where only one exchange per connection is possible.

We have seen in subsection 11.5.1, that Willer and Skvoretz (1999) needed to adjust Markovsky et al.'s (1993) original GPI-R measure to cover for networks in which positions differ with respect to the allowed number of exchanges. They do so by

re-defining (and introducing a case distinction for) the resistance equation. The precise specification of the resistance equation depends on Willer and Skvoretz' (1999) categorization of network connections which draws on the ranking and values of the node-specific parameters N, M and Q. However, as Table 11.2 shows, there are some networks (for which $N > M \geq Q > 1$) where still no profit point predictions are possible. The application to such, so called, inclusive-exclusive and inclusive-null connections requires – in addition to the updated resistance equations – the adjustment of the original Graph-theoretic power index (GPI; Markovsky, Willer, and Patton 1988) via the "correction index" 1/Q (see for details Willer and Skvoretz 1999:

Table 11.2: Observed and predicted dyadic profit splits in complex diamond, H- and T-shape structures.

Network	Match	Tie[a]	NCB	GPI-I-R/GPI-RD	Observed (s.e.)
Diamond	A:B	pr	17.01	$Profit_A = Profit_A$	18.05 (0.23)
H-2-3	C:A	pr	18.88†	$Profit_{C:A} < Profit_{C:D}$	19.39 (0.36)
	C:D	mr	20.77†	$Profit_{C:D} > Profit_{C:A}$	20.69 (0.48)
	D:B	pnr	8.38†	$Profit_{D:B} > Profit_{D:C}$	8.79 (0.31)
H-2	B:A	pr	15.28†	14.60	15.50 (0.41)
Triple-T	B:A	pnr	7.81	12.00	13.53 (0.45)
	B:C	mnr	3.94	2.20	6.12 (1.01)
	C:D	pr	14.64	16.00†	17.72 (0.93)
Double-T[b]	B:A	pr	20.66	19.60	20.67 (0.49)
	C:D	pr	12.00	12.00	12.86 (0.70)
AD[c]			2.01 (1.41)	1.67	
MD[c]			1.15 (0.70)	0.81	

Note: Observed profit splits as reported in Willer and Skvoretz (1993) and Skvoretz and Willer (1999). NCB = Network Control Bargaining model. Predictions and observations are for profit points (out of a cake of 24 profit points) of the first named structural position. GPI-I-R makes no point predictions but only compares for (in)equality of profit splits (DIAMOND and H-2-3) while GPI-RD allows for point predictions in the remaining networks. However, for GPI-RD, assumptions have to be made with whom a position with more than one possible cake to be divided exchanges first and the number of ties of such an actor are adjusted accordingly (see Lovaglia et al. 1995: 142–143).

[a]Classification of relations: pr = pure rival; mr = mixed rival; pnr = pure non-rival; mnr = mixed non-rival.

[b]The Double-T network is not stable in the sense that the negotiation and exchange structure differ (see text for more details).

[c]AD = Absolute Deviation (the sum of absolute distances between observed and predicted profit points relative to number of comparisons); MD = Mean Deviation (the Euclidean distance between observed and predicted profit points relative to number of comparisons). Figures in brackets for NCB refer to the comparison of ties of the latter three networks only.

†Daggers indicate that predicted values fall within the 95% confidence interval of the observed values. Put differently, these predictions fit the observations at the $p < 0.05$ significance level (two-tailed tests).

215–216). Willer and Skvoretz apply this GPI-I-R model to the Diamond and H-2-3 structures depicted in Figure 11.2. However, no point predictions for profit splits but only relative comparisons are reported (see the first two networks in Table 11.2).

Since we wish to confront NCB predictions with profit point predictions from other theories, we rely on experimental results from H-2, Triple-T, and Double-T (see Figure 11.2) reported in Skvoretz and Willer (1993) and the respective theoretical predictions based on Lovaglia et al.'s (1995) GPI-RD measure. It makes use of a revised interpretation of the degree index, which is part of GPI-RD. Calculation of the degree index in the original approach rests on the relative number of ties negotiating actors possess. However, if a position is allowed to exchange (more than) twice per round, its effective number of ties then depends, by assumption, on the sequence of exchanges the position engages in. To account for this 'path dependency', a position's effective number of ties need to be calculated as a weighted average of the number of exchange possibilities a position has for each of its allowed exchanges (for details see Lovaglia et al. 1995). While GPI-based measures thus can be adapted to embrace situations which deviate from the one-exchange rule, it is unclear how they could account for negotiation structures with heterogeneous cake sizes (see subsection 11.5.3).

Table 11.2 summarizes the empirical results for the networks depicted in Figure 11.2 and in addition reports on the classification of relations for each exchange tie under consideration. First, take a look at the results for the Diamond and H-2-3 structures. GPI-I-R makes no profit point predictions but allows for a ranking comparison of profit splits in different dyads. As can be seen, GPI-I-R predicts the ordinal ranking of profit splits correctly for the H-2-3 networks. Since the As in the Diamond structure are structurally equivalent, it is no surprise that GPI-I-R predicts the same profit shares for the A positions in match with a respective B. NCB calculations, on the other hand, allow for point predictions in both networks which are extremely well in line with empirical observations. For the H-2-3 structure, they all fall within the 95% confidential interval of the observed values. And, it can clearly be seen that position D in the H-2-3 structure, even though, structurally well positioned, is exploitable because he must conclude one exchange with each of his partners. For the Diamond structure, NCB predicts correctly that the As – even they are forced to conclude two exchanges – are in a far better network position than the Bs. All ties are pure rival ones and the fact that the Bs need to exchange twice only marginally deteriorates their advantageous structural position.

Now take a look at the H-2 structure where the B positions are forced to exchange twice in each round. It can be expected that exchange between B_1 and B_2 would result in an equal split of the cake due to their structural equivalence. We know that the simple H-Shape structure with a one-exchange rule (i.e., the Bs can only exchange once in every round) breaks. The bargaining network does not coincidence with the final exchange network since the Bs permanently refrain from exchange with each other and prefer a monopoly position in a 3-Line to exchange with their peripheral As (see Braun and Gautschi 2007). The 3-Line generates more profit for a rational and profit point maximizing B (almost 84% of the cake in

exchange with A) than they could earn in the H-Shape structure (only about 78% of the cake). The question now arises whether H-2 will show the same property as its simple brother? Since expected profits in pure rival relations decrease in the number of necessary exchanges, B can surely not expect 84% of the cake in exchange with A in a 3-Line where he must exchange with both As in each round. NCB predicts, for this simple structure, that B could only harvest a little more than 37% of the cake. However, this is far less than he gets from the As when he at least occasionally exchanges with the other B as well in a H-2 structure. Position B then receives about 64% of the cake in relation with A (see fifth row in Table 11.2). The H-2 structure therefore is stable in the sense that the negotiation and exchange networks coincide. And, as can be seen, NCB profit point predictions fall within the 95% confidence interval of the observed values.

We now turn to the Double-T structure where all actors need to complete two exchanges in each round.[44] Just having in short discussed the logic of network breaks, it is easy understood that the Double-T structure is not stable. The negotiation structure decays into a Double-3-Line and a Double-Dyad due to B's decision to permanently refrain from exchange with C. B can increase his profit in relation with A from about 82% of the cake in the Double-T to about 86% of the cake in the Double-3-Line by excluding C from bargaining and exchange.[45] The latter is thus forced into a dyad with D where profit points are split evenly (due to structural equivalence). As Table 11.2 shows, NCB prediction for the Double-T are most accurately in line with laboratory observations.

Finally, let us discuss the Triple-T network. Table 11.2 clearly shows that none of the NCB predictions fall within the 95% confidence interval. They are far away from observed values, especially for the B:A and B:C relation. To explain this lack of conformity with observed profit splits, take a closer look at the B:A relation. Why should B get about 56% of the cake in relation with A, as experimental findings suggest? B is in a similar situation as in the Branch333 structure (see Figure 11.1) where the observed profit for B in match with A was 7.96 or just about 33% of the whole cake (see Table 11.1). The crucial argument in favor of NCB predictions now is the fact that in the Triple-T structure, position B even depends on the A's somewhat more since C no longer is peripheral (as in the branch structure) but itself has a peripheral exchange partner D. This should increase C's negotiation power over B as well as the As negotiation power over B. It can thus be expected that B in relation

44 Note that peripheral actors (i.e., A_1, A_2, and D) are not punished if they cannot realize two exchanges in each round. They are dependent on B and C, respectively, for negotiation and exchange and those positions dictate whether peripheral actors will indeed realize two exchanges in each round. However, B and C must complete two exchanges in each round.

45 Only 4 out of 307 exchanges occurred in the B:C relation. A one-tailed exact binomial test for proportions of exchange shows that this is significantly less than could be expected if positions choose each other at random.

with A in the Triple-T structure would earn (a little) less than B in relation with A in the Branch333 structure. This is exactly what NCB predicts for the B:A relation: 7.81 profit points in favor of B. This is less than the observed 7.96 profit points of B in relation with A in the Branch333 network (see Table 11.1). The observed value of 13.53 profit points in the B:A relation therefore seems puzzling. Personal communication with David Willer, however, confirms our skepticism. The Triple-T experiments were subject to 10 rounds of negotiation and exchange only as well as a small *N*. The empirically observed values for this network, as reported in Table 11.2 should therefore be taken with care.

In sum, one can conclude that NCB predictions for networks where multiple exchanges per round are possible are more than satisfying. We now turn to a final set of networks which again, on the one hand, were subject to a one-exchange rule but then varied the value of the negotiation ties (i.e., varying cake sizes) instead.

11.5.3 Structures with different cake sizes

The final set of experiments to be discussed stems from Bonacich and Friedkin (1998) and concerns network structures with unequally valued pure rival relations (heterogeneity with respect to cake sizes). Figure 11.3 depicts the four network structures. Besides the labels for the positions, we now also label the negotiation ties where the respective value of each negotiation tie can be read off the third column in Table 11.3. For instance, there were three different 4-Line structures studied in the experiment. First, one with equal profit points of value 8 in all ties. Second, a 4-Line with cake size 6 in the two B:A relations and a cake size of 12 in the A:A relation. And third, a structure where the B:A relations negotiated over a cake size of 4 and the A:A relation haggled over a cake size of 16.

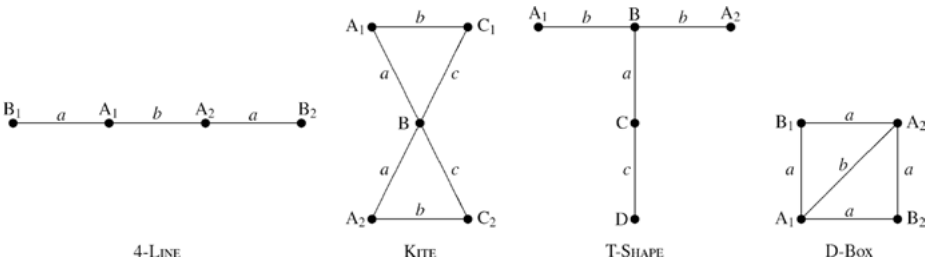


Figure 11.3: Examples of complex structures with varying cake sizes according to Bonacich and Friedkin (1988).

Unfortunately, Bonacich and Friedkin (1998) do not provide point predictions for the experimentally observed profit splits. They do, however, represent the results in

Table 11.3: Observed and Predicted Dyadic Profit Splits in Complex Structures with Varying Cake Sizes.

		Cake Size						
Network	**Match**	**a**	**b**	**c**	**NCB**	**EV**	**PD**	**Observed**
4-Line	B:A	8	8	–	2.70	3.04	2.72	4.00
		6	12	–	1.43	2.04	0.00	2.52
		4	16	–	0.60	1.36	0.00	1.48
Kite	B:A	8	14	2	2.10	1.73	0.00	3.30
		8	8	8	4.62	3.47	3.47	3.60
		8	2	14	7.04	5.33	8.00	5.07
T-Shape	B:C	9	6	9	5.04	4.50	3.00	4.50
		12	6	6	7.68	6.00	6.00	6.00
		15	6	3	11.10	7.50	9.00	7.50
D-Box	B:A	8	8	–	3.46	3.36	2.00	4.00
		6	12	–	2.09	2.16	0.00	2.73
		4	16	–	0.96	1.34	0.00	1.93
AD[b]					1.27	0.43	1.76	
MD[b]					0.46	0.18	0.58	

Note: Observed profit splits as graphically depicted in Bonacich and Friedkin (1998).
NCB = Network Control Bargaining model. Predictions and observations are for profit points of the structural position B in the relation with value *a* (cake size). Observed profit points and predictions for the Expected Value theory (EV; Friedkin 1993, 1995) and the Power-Dependence theory (PD; Yamagishi and Cook 1992) are read off Figure 2 in Bonacich and Friedkin (1998: 169) since no table with accurate profit splits is available.
[b]AD = Absolute Deviation (the sum of absolute distances between observed and predicted profit points relative to number of comparisons); MD = Mean Deviation (the Euclidean distance between observed and predicted profit points relative to number of comparisons).

graphs (p. 169, Figure 4). The observed profit points reported in Table 11.3 thus are the respective values read off these graphs as accurately as possible. This is also true for the reported predictions of the Power-Dependence theory (Yamagishi and Cook 1992) and the Expected-Value theory (Friedkin 1993, 1995). Bonacich and Friedkin (1998: 162–163) use predictions based on a slight extension of the Power-Dependence theory which then is, under certain assumptions, identical to the kernel. No adjustments have to be made to the Expected-Value theory. Bonacich and Friedkin (1998: 169) also "report" profit point predictions by the core. However, since the core is empty for the Kite network and otherwise only makes range predictions, we refrain from reporting and comparing its predictions in detail. It is sufficient to report that NCB predictions and observations for the 4-Line, the T-Shape, and the D-Box fall within the range of the prediction of the core (viz., a share of 0 to 0.5 of the respective cake sizes in the 4-Line and the D-Box, and a share of 0.4 to 0.8 of the whole cake in the T-Shape).

In networks with equally valued relations, absolute and relative profit shares of *i* in match with *j* provide identical information. However, in networks with unequally valued relations, absolute shares can become more important since a small fraction of a large cake can still outperform a large fraction of a small cake. It is therefore to be expected that *i*'s share in match with *j*, ceteris paribus, increases when the value of relation in the *i*:*k* match increases. Since the latter then becomes more favorable to *i*, *j*'s only chance to be considered for further exchange is to increase his offer to *i*.

Unfortunately, this scenario is not implemented in a "pure form" in the networks under study. However, this mechanism can be observed in the 4-Line. The value of the A:A tie was increased in steps from 8 to 16 while, however, the value of the B:A tie was changed as well. Therefore, we also consider B's relative profit shares in match with A and not just his absolute profit shares. As Table 11.3 shows, B's share of the cake in match with A decreases in absolute and relative terms when the A:A relation increases in value. B gets about 34% (i.e., 2.70/8) of the B:A cake in an equally valued 4-Line, still about 24% (i.e., 1.43/6) when the A:A tie is increased in value from 8 to 12. However, if the A:A value is further increased to 16, B only receives 15% (i.e., 0.60/4) of the cake in match with A anymore.

While NCB predictions are less accurate than predictions by the Expected Value theory (see goodness-of-fit measures reported in the last two rows of Table 11.3), they are far better than the Power Dependence predictions. And, all theories correctly predict the ordinal ranking of profit splits in the respective networks. Again, recall that observed profit splits as well as those reported for the competing theories have been read off a graphical representation in Bonacich and Friedkin (1998: Table 11.4). These figures are thus subject to some margin of error.

Take a look at the T-Shape structure in Figure 11.3. We have already discussed variants of this structure in the previous subsection. We are interested in the profit splits between B and C (value of relation a). Observations for all three variants of this structure is an equal split of the cake. Expected Value theory predicts precisely this. However, from a structural point of view, this result is not plausible. The structural position B is always more powerful than C. B has, in comparison to C, two additional peripheral exchange partners (A_1 and A_2) who are fully dependent on him for negotiation and exchange. C, on the other hand, only has one additional peripheral partner D. Since bargaining power is a function of the network embeddedness (structural independence in pure rival relations) and B is better embedded than C, his profit in match with C should exceed half of the pie. B's advantageous structural position is even reinforced by the heterogeneous values of relations – the B:A cake remains constant at $b = 6$ between experiments while profit points in the C:D relation (value *c*) decrease.

In contrast to a T-Shape with equally valued relations which breaks (i.e., no coincidence between the negotiation and exchange structures), NCB now predicts exchanges between B and C in all T-Shape structures under investigation. To see this,

consider the T-Shape where a = 9, b = 6 and c = 9. B receives 4.78 profit points (of a total of 6) in match with A (not reported in Table 11.3). In an equally valued 3-Line with cake sizes b = 6, however, he would receive 4.97 profit points out of 6 (i.e., p_{BA} = 0.8276, B's relative bargaining profit, times 6). This is more than B gets from A in the T-Shape. The necessary condition for a network break would thus be fulfilled. But since B gets even more from C, namely 5.05 out of 9, he will keep exchanging with both A and C. B thus he has no rational to refrain from permanent exchange with C (as is the case in the simple T-Shape with equally valued relations). And, as discussed in the previous paragraph, B's profit in match with C exceeds half of the pie – he gets 56% of the pie from C.

That the T-Shape is stable thus crucially hinges on the fact that the B:C relation is valuable to B in absolute terms. Even though in relative terms, B gets less from C if compared to the equally valued T-Shape. In the latter, p_{BC} = 0.6397 while in the unequally valued T-Shape with a = 9, b = 6 and c = 9, p_{BC} = 0.5605 only. However, since the absolute share of the B:C cake exceeds the absolute share of the B:A tie, B has no incentive to refrain from exchange with C.[46] The B:C relation should now proof even more valuable when it increases in value. Table 11.3 clearly shows an increase in absolute profit points for B in match with C. And, since the value of the B:A relations (i.e., b = 6) remains constant, it pays less and less for B to break into a 3-Line. Note that since the value of the B:C tie increases while (i) the B:A remains constant and (ii) the value of the C:D tie decreases, both B and C should become more powerful in match with A and D, respectively (see discussion above). But because of (ii), C gets more dependent on B for a significant absolute share of the exchange profit. Consequently, B's relative power over C also increases from p_{BC} = 0.5605 in the first T-Shape structure to p_{BC} = 0.6397 when a = 12 and, finally, to p_{BC} = 0.7372 when a = 15.

As mentioned earlier, figures from Bonacich and Friedkin (1998) should be looked at with some reservation. That the observed profit splits for the T-Shape networks are not especially plausible should increase this skepticism. If we consider the other three networks, NCB predictions could be called reasonably good. Unfortunately, Bonacich and Friedkin (1998) is the only paper which reports observed profit splits from experiments with unequally valued relations. Even though Molm, Peterson, and Takahashi (2001) also provide results on network bargaining in two unequally valued networks, the drawback of their experimental setting was the fact, that they did not put subjects into an alternating offers bargaining setting but only played a five-stage Ultimatum Game. Knowing that NCB predictions assume alternating offers bargaining, we nevertheless in short confront our predictions with their experimental findings.

46 If the value of the B:C relation is 8 or even less (everything else being constant), this logic would no longer hold. In absolute terms, B's share in match with C would then be worse than his share in match with either A. Moreover, B's share in match with either A in the respective 3-Line would outperform his B:A gain in the T-Shape. The T-Shape negotiation structure where $a \leq 8$, $b = 6$ and $c = 9$ would therefore break.

Molm, Peterson, and Takahashi (2001) considered two bargaining networks. First, a Triangle with the B:A relation of constant value a = 16, the A:C relation with constant value b = 4, and the B:C relation with varying cake sizes (c_1 = 10, c_2 = 16, and c_3 = 22). They were interested in changes in power of B over A. Since the value of the B:C tie increases, both B and C should become more powerful in relation to A. Therefore, B's profit in relation with A should increase. NCB correctly predicts an increase in B's profit over A when the B:C relation increases in value, starting with 10.28 profit points, then 11.54 profit points, and finally 12.32 profit points (always out of a total of 16 profit points). Experimentally observed values were, respectively, 9.44 profit points, 11.68 profit points, and 12.80 profit points. These values, however, have a rather large standard error such that all NCB predictions fall within the 95% confidence interval.

The second experiment was conducted using a Box structure (i.e., four actors residing on each corner of a quadrangle). B_1 is tied to B_2 (with varying cake sizes of again c_1 = 10, c_2 = 16, and c_3 = 22), B_2 is further connected via a relation of a constant value of a = 16 profit points to A_2 while A_2 in turn is tied via a relation of a constant value of b = 4 profit points to A_1. Finally, A_1 "closes the circle" with a relation of a constant value of a = 16 profit points to B_1. The line of reasoning is again the same as in the Triangle network. As the B_1:B_2 relation increases in value, B_1 (B_2) should become more powerful in relation to A_1 (A_2). Again, NCB correctly predicts the increase in absolute cake sizes. The predicted values of 9.13 profit points, 10.10 profit points, and 10.88 profit points, again out of a total of 16 profit points, fall within the 95% confidence interval of the observed values of 7.84 profit points, 9.12 profit points, and 10.40 profit points, respectively. Again, relatively large standard errors characterize the observed values.

Whether these large standard errors are due to only five rounds of offers and counteroffers or yet other aspects of the experiment cannot be answered. However, it stands out that NCB predictions are consistently larger than the observed values. It would be interesting to see whether these experimental observations would tend toward the Nash equilibrium if a true alternating offers bargaining situation had been implemented. In general, observations on the development of bilateral negotiations (i.e., offers and counteroffers) until an agreement is reached would show whether experimental subjects ever reach the Nash equilibrium and if so, how fast this occurs. Adjusting exchange theories to be able to make predictions about the evolution of offers and counteroffers over time – and not just about the profit split in the exchange equilibrium – would thus be a desirable next step.

11.6 Conclusion

This paper presented an approach to the study of complex bargaining and exchange structures. Partly due to experimental evidence on negatively connected

exchange networks with a one- exchange rule, model building has widely neglected possibly more complex features of exchange networks. Therefore, only a few theories are capable of making profit point predictions for bargaining and exchange structures where (i) the one-exchange rule is relaxed and positions are allowed to conclude more than one exchange per round, and, (ii) with unequally valued relations across the bargaining structure. However, there is no theory which can be applied to bargaining and exchange structures which are characterized by (i) and (ii). The NCB model seeks to close this gap.

In the course of developing the new model, we have found a weakness in the well-established network classification of negatively and positively connected relations. We therefore introduce a new and more precise classification of network relations which embrace negatively and positively connected ties as special cases. The new network classification is based on structural features and two node specific, but endogenous network parameters: the number of ego's bargaining partners, the number of ego's bargaining relations, and the number of exchanges ego intends to complete. The resulting parsimonious classification allows for unique profit point predictions in networks with combinations of different relational aspects.

In accordance with other theories, our approach reflects the idea that rational actors take advantage of their structural positions in negotiations. Moreover, positions are now aware that the number of exchanges which must be concluded and the value of the respective bargaining ties (i.e., the size of the cakes to be divided in bilateral bargaining) can further deteriorate or improve the (dis)advantage of a specific network position. Contrary to other sociological models, we take into account that negotiation partners pursue their self-interest and thus specify the actors' optimization problem and show where these network parameters enter the choice calculus and how they influence decision making. To comply with this demand we have combined the generalized Nash bargaining solution from game theory with the assumption that both relational features and network positions affect exchange outcomes. The applications section has shown that the resulting Network Control Bargaining (NCB) model makes predictions which closely correspond to experimental results by Bonacich and Friedkin (1998), Molm, Peterson, and Takahashi (2001), Skvoretz and Willer (1993), and Willer and Skvoretz (1999).

While the presented NCB approach can handle different relational features of exchange networks, it still assumes interindividual heterogeneity with respect to characteristics such as age, gender, education or wealth (status). Since it is known that such attributes can matter for negotiation results (e.g., D'Exelle et al. 2010; Dufwenberg and Muren 2006; Eckel and Grossman 2001; Holm and Engsfeld 2005; Schwieren and Sutter 2008; Solnick 2001), a broader approach should account for non-network related characteristics which may influence bargaining outcomes. Such personal variables, a_{ij}, could be taken into account via the actors' individual bargaining powers: $b_i = f(z_i, c_i; a_{ij})$.

Besides these limitations which prevent that the model captures all relevant aspects of real world negotiations, it must be emphasized that the approach rests on strong premises. They ensure that the theoretical model mirrors the bargaining protocol of laboratory exchange networks. However, conditions in laboratory experiments should not be set in stone. It thus would be a worthwhile task to construct less artificial (laboratory) experiments while concurrently broadening the theoretical models for sociological exchange studies.

The behavioral postulate of the model would mark a good point of departure for such a task. According to our theory, each actor is the neoclassical selfish profit maximizer. Even though the experimental protocol tries as good as possible to induce profit point maximizing behavior, it cannot completely suppress additional motivations such as fairness and aversion to inequality. These motivations have been studied widely in the recent literature on behavioral game theory (for an overview and introductions, see Camerer 2003) and shown to be present (e.g., Dictator Game, Ultimatum Game) even in best controlled and artificial laboratory experiments. Likewise, experimental subjects engaging in repeated negotiation and exchange may produce positive feelings towards their exchange partners. They form attachments and make commitments in durable negotiations (e.g., Lawler and Yoon 1996). While some experimental protocols try to inhibit such behavior by rotating subjects through all positions of a network, it only helps to forfend artificial laboratory conditions. An increase in relational cohesion stabilizes exchange relations, breeds trust and is to the benefit of both bargaining partners. These aspects of durable exchange relations deserve attention because of their importance in everyday life.

It is well established through experiments that motivations such as fairness and aversion to inequality unfold equal forces even across different cultural backgrounds (e.g., Camerer 2003; Henrich 2000). Nevertheless, there seem to be norms in distributive bargaining which could further affect outcomes in real world settings (Gautschi 2018). As Young (1996: 116) puts it: "[. . .] norms condition the parties to expect certain outcomes that depend on the bargaining context." Moreover, such norms or convention can even change over time. A realistic model of bargaining and exchange should thus be able to account for the bargaining context. The real world is not the clean laboratory environment with its legitimate tendency to create congeneric situations.

In sum, while the current model is broader in scope and thus capable of making predictions for more complex bargaining and exchange networks, it nevertheless lacks some relevant real world features, as just discussed above. Therefore, the generalization of the current model beyond relational features (such as more than one exchange per round) and unequally valued relations is important and worthwhile. Especially if the model should be suited for the analysis of real bargaining and exchange situations. They are hardly ever characterized by the artificial and highly controlled conditions of a laboratory experiment.

References

Aumann, Robert J. and Roger B. Myerson 1988. "Endogenous Formation of Links between Players and Coalitions: An Application of the Shapley Value." Pp. 175–191 in *The Shapley Value. Essays in Honor of Lloyd S. Shapley*, ed. Alvin E. Roth. Cambridge: Cambridge University Press.

Bala, Venkatesh and Sanjeev Goyal. 2000. "A Noncooperative Model of Network Formation." Econometrica 68: 1181–1229.

Bayati, Mohsen, Christian Borgs, Jennifer Chayes, Yash Kanoria, and Andrea Montanari. 2015. "Bargaining Dynamics in Exchange Networks." *Journal of Economic Theory* 156: 417–454.

Bienenstock, Elisa Jayne and Phillip Bonacich. 1992. "The Core as a Solution to Negatively Connected Exchange Networks." *Social Networks* 14: 231–243.

Bienenstock, Elisa Jayne and Phillip Bonacich. 1993. "Game Theory Models for Exchange Networks: Experimental Results." *Sociological Perspectives* 36: 117–135.

Bienenstock, Elisa Jayne and Phillip Bonacich. 1997. "Network Exchange as a Cooperative Game." *Rationality and Society* 9: 37–65.

Binmore, Kenneth G. 1985. "Bargaining and Coalitions." Pp. 269–304 in *Game-theoretic Models of Bargaining*, ed. Alvin E. Roth. Cambridge: Cambridge University Press.

Binmore, Kenneth G. 1987. "Nash Bargaining Theory II." Pp. 61–76 in *The Economics of Bargaining*, eds. by Kenneth G. Binmore and Partha Dasgupta. Oxford: Blackwell.

Binmore, Kenneth G. 1992. *Fun and Games: A Text on Game Theory*. Lexington, Mass.: D.C. Heath and Company.

Binmore, Kenneth G. 1998. *Game Theory and the Social Contract, Vol. 2: Just Playing*. Cambridge: MIT Press.

Bonacich, Phillip. 1998. "A Behavioral Foundation for a Structural Theory of Power in Exchange Networks." *Social Psychology Quarterly* 61: 185–198.

Bonacich, Phillip. 1999. "An Algebraic Theory of Strong Power in Negatively Connected Networks."*Journal of Mathematical Sociology* 23: 203–224.

Bonacich, Phillip and Elisa Jayne Bienenstock. 1995. "When Rationality Fails: Unstable Exchange Networks with Empty Cores." *Rationality and Society* 7: 293–320.

Bonacich, Phillip and Noah E. Friedkin. 1998. "Unequally Valued Exchange Relations." *Social Psychology Quarterly* 61: 160–171.

Box, G. and D. Cox. 1964. "An Analysis of Transformations." *Journal of the Royal Statistical Society* Series B 26: 211–264.

Braun, N. and T. Gautschi. 2006. "A Nash Bargaining Model for Simple Exchange Networks." *Social Networks* 28: 1–23.

Braun, N. and T. Gautschi. 2007. "Who Exchanges With Whom? A Theory of Exchange Ties and Its Application to Simple Networks." Working Paper, University of Munich and Bern.

Burke, Peter J. 1997. "An Identity Model for Network Exchange." *American Sociological Review* 62: 134–150.

Camerer, Colin F. 2003. *Behavioral Game Theory. Experiments in Strategic Interaction*. Princeton: Princeton University Press.

Chakraborty, Tanmoy and Michael Kearns. 2008. "Bargaining Solutions in a Social Network." Pp. 548–555 in *4th International Workshop on Internet and Network Economics*, eds. Christos Papadimitriou and Shuzhong Zhang. Berlin: Springer.

Chakraborty, Tanmoy, Michael Kearns, and Sanjeev Khanna. 2009. "Network Bargaining: Algorithms and Structural Results.". Pp 159–168 in *Proceedings of the 10th ACM Conference on Electronic Commerce*, eds. John Chuang, Lance Fortnow, and Pearl Pu. Stanford, CA: ACM.

Coleman, James S. 1973. *The Mathematics of Collective Action*. Chicago: Aldine.

Coleman, James S. 1990. *Foundations of Social Theory*. Cambridge: The Belknap Press of Harvard University Press.
Cook, Karen S. and Richard M. Emerson. 1978. "Power, Equity, and Commitment in Exchange Net works." *American Sociological Review* 43: 721–739.
Cook, Karen S., Richard M. Emerson, Mary R. Gillmore and Toshio Yamagishi. 1983. "The Distribution of Power in Exchange Networks: Theory and Experimental Results." *American Journal of Sociology* 89: 275–305.
Demange, Gabrielle and Myrn H. Wooders. (eds.) 2005. *Group Formation in Economics: Networks Clubs, and Coalitions*. Cambridge: Cambridge University Press.,
D'Exelle, Ben, Els Lecoutere, and Björn V. Campenhout. 2010. "Social Status and Bargaining when Resources are Scarce: Evidence from a Field Lab Experiment." Working Paper Series, University of East Anglia, Centre for Behavioural and Experimental Social Science (CBESS) 10-09, School of Economics, University of East Anglia, Norwich, UK.
Dutta, Bhaskar and Matthew O. Jackson. (eds.) 2003. *Networks and Groups: Models of Strategic Formation*. Berlin: Springer.
Dufwenberg, Martin and Astri Muren. 2006. "Generosity, Anonymity, Gender." *Journal of Economic Behavior & Organization*. 61: 42–49.
Eckel, Catherine and Philip Grossman. 2001. "Chivalry and Solidarity in Ultimatum Games." *Economic Inquiry* 39: 171–188.
Emerson, Richard M. 1972. "Exchange Theory, Part II: Exchange Relations and Networks." Pp. 58–87 in *Sociological Theories in Progress*, Vol. 2, eds. Joseph Berger, Morris Zelditch, and Bo Anderson. Boston: Houghton-Mifflin.
Emerson, Richard M. 1981. "Social Exchange Theory." Pp. 30–65 in *Social Psychology: Sociological Perspectives*, eds Morris Rosenberg and Ralph H. Turner. New York: Basic Books.
Feld, Scott L. 1991. "Why Your Friends Have More Friends than You Do." *American Journal of Sociology* 96: 1464–1477.
Foster, Dean and H. Peyton Young. 1990. "Stochastic Evolutionary Game Dynamics." *Theoretical Population Biology* 38: 219–232.
Friedkin, Noah E. 1986. "A Formal Theory of Social Power." *Journal of Mathematical Sociology* 12: 103–126.
Friedkin, Noah E. 1992. "An Expected Value Model of Social Power: Predictions for Selected Exchange Networks." *Social Networks* 14: 213–229.
Friedkin, Noah E. 1993. "An Expected Value Model of Social Exchange Outcomes." Pp. 163–193 in *Advances in Group Processes*, Vol. 10, eds by Edward J. Lawler, Barry Markovsky, Karen Heimer and Jodi O'Brien. Greenwich, CT: JAI Press.
Friedkin, Noah E. 1995. "The Incidence of Exchange Networks." *Social Psychology Quarterly* 58: 213–221.
Gautschi, Thomas. 2002. *Trust and Exchange. Effects of Temporal Embeddedness and Network Embeddedness on Providing and Dividing a Surplus*. Amsterdam: Thela Thesis.
Gautschi, Thomas. 2018. "Risk Aversion and Network Exchange. Experiments on Network Exchange and Heterogeneous Risk Preferences." Working Paper. University of Mannheim.
Gould, Roger V. 2002. "The Origins of Status Hierarchies: A Formal Theory and Empirical Test." *American Journal of Sociology* 107: 1143–1178.
Heckathorn, Douglas D. 1980. "A Unified Model for Bargaining and Conflict." *Behavioral Science* 25: 261–284.
Henrich, Joseph. 2000. "Does Culture Matter in Economic Behavior. Ultimatum Game Bargaining among the Machiguenga of the Peruvian Amazon." *American Economic Review* 90: 973–979.

Holm, H. and P. Engsfeld. 2005. "Choosing Bargaining Partners: An Experimental Study on the Impact of Information About Income, Status and Gender." *Experimental Economics* 8: 183–216.

Jackson, Matthew O. and Asher Wolinsky. 1996. "A Strategic Model of Social and Economic Networks." *Journal of Economic Theory* 71: 44–74.

Kalai, Ehud and Meir Smorodinsky. 1975. "Other Solutions to Nash's Bargaining Problem." *Econometrica* 43: 513–518.

Kleinberg, Jon and Éva Tardos. 2008. "Balanced Outcomes in Social Exchange Networks." Pp. 295–304 in *Proceedings of the 40th Annual ACM Symposium on Theory of Computing*, New York: ACM.

Lawler, Edward J. and Jeongkoo Yoon. 1996. "Commitment in Exchange Relations: Test of a Theory of Relational Cohesion." *American Sociological Review* 61: 89–108.

Lovaglia, Michael J., John Skvoretz, David Willer and Barry Markovsky. 1995. "Negotiated Exchanges in Social Networks." *Social Forces* 74: 123–155.

Malinowski, Bronislaw. 1922. *Argonauts of Western Pacific. An Account of Native Enterprise and Adventure in the Archipelagoes of Melanesian New Guinea*. London: Routledge.

Markovsky, Barry, David Willer and Travis Patton. 1988. "Power Relations in Exchange Networks." *American Sociological Review* 53: 220–236.

Markovsky, Barry, David Willer and Travis Patton. 1990. "Theory, Evidence and Intuition" (Reply to Yamagishi and Cook). *American Sociological Review* 55: 300–305.

Markovsky, Barry, John Skvoretz, David Willer, Michael J. Lovaglia and Jeffrey Erger. 1993. "The Seeds of Weak Power: An Extension of Network Exchange Theory." *American Sociological Review* 58: 197–209.

Markovsky, Barry, David Willer, Brent Simpson and Michael J. Lovaglia. 1997. "Power in Exchange Networks: Critique of a New Theory" (Comment on Yamaguchi 1996). *American Sociological Review* 62: 833–837.

Molm, Linda D., Gretchen Peterson, and Nobuyuki Takahashi 2001. "The Value of Exchange." *Social Forces* 80: 159–184.

Moulin, Hervé. 1984. "Implementing the Kalai-Smorodinsky Bargaining Solution." *Journal of Economic Theory* 33: 32–45.

Muthoo, Abhinay. 1999. *Bargaining Theory with Applications*. Cambridge: Cambridge University Press.

Nash, John F. 1950. "The Bargaining Problem." *Econometrica* 18: 155–162.

Nash, John F. 1951. "Non-Cooperative Games." *Annals of Mathematics* 54: 286–295.

Nash, John F. 1953. "Two-Person Cooperative Games." *Econometrica* 21: 128–140.

Osborne, Martin J. and Ariel Rubinstein. 1990. *Bargaining and Markets*. San Diego: Academic Press.

Perry, M. and P.J. Reny. 1993. "A Non-cooperative Bargaining Model with Strategically Timed Offers." *Journal of Economic Theory* 59: 50–77.

Rubinstein, Ariel. 1982. "Perfect Equilibrium in a Bargaining Model." *Econometrica* 50: 97–109.

Schwieren, Christiane and Matthias Sutter. 2008. "Trust in Cooperation or Ability? An Experimental Study on Gender Differences." *Economics Letters* 99: 494–497.

Simpson, Brent and David Willer. 1999. "A New Method for Finding Power Structures." Pp. 270–284 in *Network Exchange Theory*, ed. David Willer. Westport, CT: Praeger.

Skvoretz, John and Thomas J. Fararo. 1992. "Power and Network Exchange: An Essay toward Theoretical Unification." *Social Networks* 14: 325–344.

Skvoretz, John and Michael J. Lovaglia. 1995. "Who Exchanges with Whom: Structural Determinants of Exchange Frequency in Negotiated Exchange Networks." *Social Psychology Quarterly* 58: 163–177.

Skvoretz, John and David Willer. 1991. "Power in Exchange Networks: Setting and Structural Variations." *Social Psychology Quarterly* 54: 224–238.
Skvoretz, John and David Willer. 1993. "Exclusion and Power: A Test of Four Theories of Power in Exchange Networks." *American Sociological Review* 58: 801–818.
Skyrms, Brian and Robin Pemantle. 2000. "A Dynamic Model of Social Network Formation." *Proceedings of the National Academy of Sciences of the United States of America* 97: 9340–9346.
Slikker, Marco and Anne van den Nouweland. 2001. *Social and Economic Networks in Cooperative Game Theory*. Boston, MA: Kluwer.
Solnick, Sara J. 2001. "Gender Differences in the Ultimatum Game." *Economic Inquiry* 39: 189–200.
Thye, Shane R., Michael J. Lovaglia and Barry Markovsky. 1997. "Responses to Social Exchange and Social Exclusion in Networks." *Social Forces* 75: 1031–1049.
Vega-Redondo, Fernando. 2007. *Complex Social Networks*. Cambridge: Cambridge University Press.
Watts, Alison. 2001. "A Dynamic Model of Network Formation." *Games and Economic Behavior* 34: 331–341.
Willer, David (ed.) 1999. *Network Exchange Theory*. Westport, CT: Praeger.
Willer, David and Pamela Emanuelson 2008. "Testing Ten Theories." *Journal of Mathematical Sociology* 32: 165–203.
Willer, David and John Skvoretz. 1999. "Network Connection and Exchange Ratios: Theory, Predictions, and Experimental Tests." Pp. 195–225 in *Network Exchange Theory*, ed. David Willer. Westport, CT: Praeger.
Yamagishi, Toshio and Karen S. Cook. 1990. "Power Relations in Exchange Networks: A Comment on 'Network Exchange Theory'." *American Sociological Review* 55: 297–300.
Yamagishi, Toshio and Karen S. Cook. 1992. "Power in Exchange Networks: A Power-Dependence Formulation." *Social Networks* 14: 245–266.
Yamagishi, Toshio, Mary R. Gillmore and Karen S. Cook. 1988. "Network Connections and the Distribution of Power in Exchange Networks." *American Journal of Sociology* 93: 833–851.
Yamaguchi, Kazuo. 1996. "Power in Networks of Substitutable and Complementary Exchange Relations: A Rational Choice Model and an Analysis of Power Centralization." *American Sociological Review* 61: 308–332.
Yamaguchi, Kazuo. 1997. "The Logic of the New Theory and Misrepresentations of the Logic" (Reply to Markovsky et al. 1997). *American Sociological Review* 62: 838–841.
Yamaguchi, Kazuo. 2000. "Power in Mixed Exchange Networks: A Rational Choice Model." *Social Networks* 22: 93–121.
Young, H. Peyton. 1993. "An Evolutionary Model of Bargaining." *Journal of Economic Theory* 59: 145–168.
Young, H. Peyton. 2001. *Individual Strategy and Social Structure. An Evolutionary Theory of Institutions* Princeton: Princeton University Press.

Part II: **Experimental Tests**

Ozan Aksoy

12 Social Identity and Social Value Orientations

Abstract: This study provides an extension of the social value orientation model and a tool, other-other Decomposed Games, to quantify the influence of social identity on social value orientations. Social identity is induced experimentally using the minimal group paradigm. Subsequently, the weights subjects add to the outcomes of outgroup others relative to ingroup others and to the absolute difference between the outcomes of ingroup and outgroup others are estimated. Results are compared to a control condition in which social identity is not induced. Results show that when the outgroup is better off than the ingroup, the average subject is spiteful: they derive negative utility from the outcomes of the outgroup other. When the ougroup is worse off than the ingroup, the average subject attaches similar weights to the outcomes of outgroup and ingroup others. There is also significant variation across subjects with respect to the level of ingroup bias.

12.1 Introduction

A quick glance at any major newspaper nowadays will, very likely, show that humans are willing to incur significant costs to members of outgroups in order to protect or better the outcomes for the ingroup. Children whose parents are caught crossing the border illegally are separated from their parents to deter illegal entry to the US.[1] Legal and "skilled" immigrants in the UK are required to pay an annual Immigration Health Surcharge of £400, in addition to the expensive visa fees and the usual tax contributions to the National Health Service.[2] Almost all countries are

1 https://www.independent.co.uk/news/world/americas/us-politics/us-immigration-children-audio-trump-border-patrol-separate-families-parents-detention-center-a8405501.html
2 https://www.gov.uk/healthcare-immigration-application

Notes: I thank John Jensenius III for his help in programming and his comments on the design. I also thank participants of the CESS Colloquium Series, Akitaka Matsuo, Wojtek Przepiorka, and an anonymous reviewer. The experiment reported in this chapter was conducted at Nuffield CESS. This research is supported by the Netherlands Organization for Scientific Research (NWO) under grant 446-13-004. The Stata code and the dataset used to produce the results reported in this chapter are available for replication at https://osf.io/wxra3/.

Ozan Aksoy, UCL Social Research Institute, University College London

https://doi.org/10.1515/9783110647495-012

imposing tariffs or quotas to certain foreign goods to protect domestic producers.[3] There are very strong barriers to international labor mobility, legal or illegal. These examples show that humans value the outcomes of ingroup members more than the outcomes of outgroup members. But how much more?

Harvard economist Dani Rodrik asks a simple and related question: "how strong a preference must we have for our fellow citizens relative to foreigners to justify the existing level of barriers on international labor mobility" (Rodrik 2017). After a simple calculation based on a plausible scenario, he concludes that "we must place a weight on the utility of fellow citizens that is at least between four and five times greater than the weight we place on foreigners". Or equivalently, a foreigner must be worth less than 22 percent of a fellow citizen. In a similar exercise, Kopczuk et al. (2005) argue that the observed levels of international assistance to developing countries imply that Americans must value their fellow citizens' outcomes about six times more than they value foreign citizens' outcomes. Or equivalently, a foreign citizen is worth 17 percent of a fellow citizen. It thus appears that the weight attached to the outcomes of outgroup members is 17 to 22 percent of the weight attached to the outcomes of ingroup members. But how accurate is this estimate? The calculations of Rodrik (2017) and Kopczuk et al. (2005) are rather indirect. They use the levels of barriers on international labor mobility and foreign assistance to estimate the relative weights. Barriers to labor mobility and foreign assistance are complex policies that are influenced by many factors, in addition to how much actors weight the outcomes of outgroups.

In this chapter, I propose a simple tool to estimate directly how much actors weight the outcomes of outgroup members relative to the outcomes of ingroup members. I build on the social value orientation and the minimal group paradigm literatures. The social value orientation literature investigates how actors value certain outcome allocations between *self* and *others* (Griesinger and Livingston 1977; McClintock 1972; Schulz and May 1989). Cooperative orientation, maximizing the sum of the payoffs for self and others; competitive orientation, maximizing the difference between the payoffs for self and others in favor of self; equality orientation, minimizing the inequality between outcomes are some of the social value orientations distinguished in the literature. Numerous methods have been developed to measure social value orientations (Aksoy and Weesie 2012; Aksoy and Weesie 2013; Kuhlman, Brown, and Theta 1992; Liebrand and McClintock 1988; Murphy, Ackermann, and Handgraaf 2011; Van Lange 1999). All of these methods involve some form of Decomposed Games in which subjects are asked to choose a certain outcome allocation between self and others among a menu of possible self-other allocations.

There is a hidden but strong link between the social value orientation literature and the minimal group paradigm. The minimal group paradigm is about how actors

3 http://tariffdata.wto.org/

value outcome allocations between *two others*, e.g., one ingroup and one outgroup (Tajfel 1970; Tajfel et al. 1971). In other words, as opposed to the social value orientation literature, the minimal group paradigm involves other-other allocations instead of self-other allocations. In fact, in minimal group experiments self-other allocations are carefully avoided. This is because in minimal group experiments, subjects' own individual interests should not be at stake to isolate the influence of mere social categorization from any form of realistic conflict (Sherif et al. 1961). Because of this omission of self from outcome allocation tasks, the tools of the social value orientation literature cannot readily be applied to minimal group settings.

In this chapter, I explicitly bridge the social value orientation literature with the minimal group paradigm. I extend the social value orientation model to other-other allocations (Macro and Weesie 2016). Moreover, I also show that the classical self-other Decomposed Games of the social value orientation literature can easily be adapted to other-other allocations, hence to minimal group setting. Using experimental data, I quantify the influence of social identity on social value orientations.

12.2 Theory: Social value orientations in other-other allocations

In the classical social value orientation model, for an outcome allocation for self (x) and other (y), an actor i attaches a w_i weight to the outcome of other such that (Aksoy and Weesie 2012; Aksoy and Weesie 2014; Griesinger and Livingston 1977; McClintock 1972):

$$U_i(x,y) \equiv U_i^*(x,y\,;\,w_i) = x + w_i y. \tag{12.1}$$

Let's now assume that there are two types of others, ingroup and outgroup. Let I (O) denote the set of ingroup (outgroup) others. Consider an other-other allocation situation in which the ingroup other gets y^I, the outgroup other gets y^O, and there is no outcome for self, i.e., $x = 0$. In this situation, the social value orientation model in (12.1) can be written as:

$$U_i\left(y^I, y^O\right) \equiv U_i^*\left(y^I, y^O;\, w_i^I, w_i^O\right) = w_i^I y^I + w_i^O y^O \tag{12.2}$$

where w_i^I and w_i^O are the weights actors attach to the outcomes of ingroup and outgroup others, respectively. Because utility is defined up to positive affine transformations, and assuming that $w_i^I > 0$, equation (12.2) can be written as:

$$U_i\left(y^I, y^O\right) \equiv U_i^*\left(y^I, y^O;\, \Theta_i^O\right) = y^I + \Theta_i^O y^O \text{ with } \Theta_i^O = \frac{w_i^O}{w_i^I}. \tag{12.3}$$

Equation (12.3) is now equivalent to the model in equation (12.1) where the outcomes for self and other are replaced by the outcomes for ingroup other and outgroup other, respectively. Consequently, the weight actors attach to the outcomes of outgroup others relative to ingroup others can easily be estimated using other-other Decomposed Games just as the social value orientations are estimated with self-other Decomposed Games (Aksoy and Weesie 2012).

Finally, social value orientation research has shown that some people also consider inequality in outcomes, such as those with equality or maximin orientations (Aksoy and Weesie 2012; Grzelak, Iwinzki, and Radzincki 1977; Macro and Weesie 2016; Schulz and May 1989). These orientations are typically captured by adding another term in equation (12.1), the absolute inequality between the outcomes for self and other. In the other-other allocation case, an equivalent term will be adding the absolute inequality between the outcomes for ingroup and outgroup others. Thus,[4]

$$U_i(y^I, y^O) \equiv U_i^*(y^I,\ y^O;\ \Theta_i^O, \beta_i) = y^I + \Theta_i^O y^O - \beta_i |y^I - y^O|. \tag{12.4}$$

A useful reinterpretation of the model in (12.4) is the following:

$$U_i(y^I, y^O) \equiv U_i^*(y^I,\ y^O;\ \Theta_i^O, \beta_i) = \begin{cases} y^I + \frac{\Theta_i^O + \beta_i}{1-\beta_i} y^O \text{ if } y^I \geq y^O \\ y^I + \frac{\Theta_i^O - \beta_i}{1+\beta_i} y^O \text{ if } y^I < y^O \end{cases} \tag{12.5}$$

Equation (12.4) is mathematically equivalent to (12.5) when $-1 < \beta < 1$. My empirical results below will indeed show that $-1 < \beta < 1$, hence I will use (12.4) and (12.5) interchangeably. The specification in (12.5) is easier to interpret than that in (12.4). In (12.5) we have two separate terms that represent the weights attached to the outcomes of the outgroup other relative to the ingroup. When the outgroup is worse off than the ingroup this weight is $\frac{\Theta_i^O + \beta_i}{1-\beta_i}$. When the outgroup other is better off than the ingroup other the weight to the outgroup is $\frac{\Theta_i^O - \beta_i}{1+\beta_i}$. Hence, while (12.5) and (12.4) are mathematically equivalent, specification (12.5) provides an alternative, more convenient interpretation.

One could in principle modify the model in (12.4) by taking not the *outcomes* for the ingroup or outgroup, but the *utilities* for the ingroup and outgroup. For example, one could define $U_i'(y^I, y^O)$ as $U_i'(y^I, y^O) = U_i^I(y^I, y^O) + \Theta_i'^O U^O(y^I, y^O) - \beta_i' |U^I(y^I, y^O) - U^O(y^I, y^O)|$. That is, replacing the outcomes of ingroup and outgroup others with

4 When inequality concerns are introduced, in other-other allocations in which self gets zero it can be argued that actors may take two additional terms into account: the difference between outcomes for ingroup others and self as well as the difference between outgroup others and self. In this case, the model can be written as $U = x + w_i^I y^I + w_i^O y^O - b_i^I |y^I - x| - b_i^O |y^O - x| - \beta_i |y^I - y^O|$. Because in other-other allocations $x = 0$, this alternative formulation can be re-arranged (assuming $y^I > 0,\ y^O > 0,\ w_i^I > b_i^I$) such that $U = y^I + \frac{w_i^O - b_i^O}{w_i^I - b_i^I} y^O - \beta_i |y^I - y^O|$, which is equivalent to the formulation in equation (12.4) with $\frac{w_i^O - b_i^O}{w_i^I - b_i^I} = \Theta_i^O$.

utilities for ingroup and outgroup others. Alternatively, one could define equality as $y^I = \Theta y^O$, and hence replace the term $|y^I - y^O|$ in (12.4) with $|y^I - \Theta y^O|$. These alternative formulations are examples of interdependent utility, i.e., actors are interested in utilities of other actors not just outcomes (Becker 1993). In this chapter, I don't consider interdependent utility.

12.3 Method

12.3.1 Subjects

186 subjects were recruited with the Online Recruitment System for Economic Experiments (ORSEE; Greiner 2004). Majority of the subjects were students at the University of Oxford from a variety of different study fields. Subjects were on average 30 years old (S.D.=14) and 58% of them were female.[5]

12.3.2 Procedure

Subjects participated in one of ten sessions in Hilary Term (February-March) 2014. Subjects in seven sessions were assigned to the experimental group and in the remaining three sessions to the control group. Seven sessions were run between 16 and 24 subjects and three sessions were run between 12 to 14 subjects. Subjects sat randomly in one of the cubicles in the Centre for Experimental Social Sciences (CESS) lab at Nuffield College, University of Oxford. Subjects could not see each other or the experimenter during the experiment. This also meant that the subjects were not fully aware of the total number of subjects in the experiment, though they might have a rough guess about the session size (the median session size was 20). The experiment was carried out on computers using z-tree (Fischbacher 2007).

12.3.2.1 Experimental group

After general instructions, subjects in the experimental group were shown five pairs of paintings by Wassily Kandinsky and Paul Klee. For each pair, subjects chose the painting they liked more. 50% of subjects in a session were classified as Kandinskys,

5 The experiment reported here is embedded in a larger study which included additional unrelated tasks. These additional tasks were administered *after* the procedure described here took place and were analyzed elsewhere. See Aksoy (2015) for details of these additional tasks.

and the remaining 50% as Klees, based on subjects' relative preferences. Each subject was privately informed about his/her group.

After classification, a collective quiz in which subjects guessed the painters of two paintings (Klee or Kandinsky) was administered. Subjects earned £0.8 if at least 50% of their group correctly guessed the two painters. Subjects earned a further £0.8 if their group correctly answered as many questions as the other group. Quiz results were shown only after the experiment was completed.

After the collective quiz, subjects made decisions in 10 other-other Decomposed Games shown in the appendix (Table A1). The order of these 10 games was varied in two factors. These 10 games were modified versions of the self-other Decomposed Games used by Aksoy and Weesie (2012).[6] Recipients in these Decomposed Games were a randomly selected ingroup member and a randomly selected outgroup member. At the end of the experiment, one Decomposed Game was selected at random, and two actual other subjects received the tokens based on a subject's decision (20 tokens = £1). Similarly, each subject was a recipient for a randomly selected other subject.

12.3.2.2 Control group

The control group followed the procedure above but without inducing group identity. Subjects stated their preferences in the same 5 painting pairs. However, they were not classified as Klees or Kandinskys. They completed the same guessing quiz but they were rewarded for their individual success: for each correct guess, a subject earned £0.8. Finally, subjects decided in the same 10 other-other Decomposed Games. Different from the experimental group, the two recipients were two other subjects randomly selected from the session, without any reference to any groups.

12.4 Results

I follow the estimation procedure described in Aksoy and Weesie (2012). In this procedure, the outcomes in an other-other Decomposed Game are transformed into utilities via equation (12.4). In addition, an additive random utility term ϵ is added to the model to have stochastic behavioral predictions. The random utility term makes the utility model statistically estimable. How much a subject i prefers option A relative to option B is the utility difference in options A and B in a game:

6 The modifications aimed to improve the statistical precision to estimate the social value orientation parameters based on the results reported in Aksoy and Weesie (2012) and additional simulations.

$$U_{AB}(y^I, y^O; \Theta_i^O,\ \beta_i) = (y_A^I - y_B^I) + \Theta_i^O (y_A^O - y_B^O) - \beta_i (|y_A^I - y_A^O| - |y_B^I - y_B^O|) + (\epsilon_A - \epsilon_B) \tag{12.6}$$

where y_A^I is the outcome for ingroup other in option A and y_B^O is the outcome for outgroup other in option B in a Decomposed Game. A subject prefers option A in a Decomposed Game when $U_{AB} > 0$. Following Aksoy and Weesie (2012), (Θ^O, β) are treated as bi-variate normally distributed variables and ϵ is assumed to have an independent normal distribution with zero mean and nonzero variance. This implies a multilevel probit model in which the dependent variable is a subject's preferences in the 10 Decomposed Games and independent variables are the outcome differences given in equation (12.6). The distribution of (Θ^O, β), the variance of $(\epsilon_A - \epsilon_B)$, and the empirical Bayes predictions (posterior means) of Θ^O and β per subject are estimated with the Stata program GLLAMM (Rabe-Hesketh, Skrondal, and Pickles 2002). The replication material with the Stata code and the data are available at https://osf.io/wxra3/.

Table 12.1 and Figure 12.1 show the results. When social identity is induced (experimental group), the estimated mean of Θ^O is 0.2 which is significantly different from both zero and one. This means that when the inequality between ingroup and outgroup is zero, the average weight subjects add to the outcomes of outgroup others is only 20% of the weight they add to the outcomes of ingroup others. The estimated mean of β is 0.39 and significantly different from zero (and one). This

Table 12.1: Social value orientation estimates for the experimental and control groups. Θ^O = "outgroup cooperative orientation parameter"; β = "equality orientation parameter"; ϵ_A, ϵ_B= evaluation error. For the variances, p-values are derived from the correct boundary tests using the mixture distribution Self and Liang (1987).

	Experimental Group		Control Group	
Parameter	**Coeff.**	**S.E.**	**Coeff.**	**S.E.**
mean(Θ^O)	0.200***	0.059	1.003***	0.154
mean(β)	0.389***	0.032	0.235**	0.094
var(Θ^O)	0.327***	0.087	0.095	0.073
var(β)	0.050***	0.020	0.063	0.062
cov(Θ^O, β)	−0.099***	0.038	−0.073	0.058
var($\epsilon_A - \epsilon_B$)	0.021***	0.002	0.021***	0.005
N(Subject)	146		40	
N(Decision)	1460		400	
log-likelihood	−723.762		−140.875	

***p 2-sided<0.001;**p 2-sided<0.01

means that while subjects add a small weight to the outcomes of outgroup others relative to ingroup others, they are still concerned with reducing inequality between ingroup and outgroup others. There is also a negative correlation between Θ^O and β.

A further noteworthy finding is the significant and large variance of Θ^O in the experimental group. An estimated variance of 0.34 implies that although on average there is significant ingroup bias, there is also a significant variation among subjects regarding the level of ingroup bias. A minority of subjects in fact have Θ^O values very close to, but never exceeding one (see Figure 12.1a). These subjects could be described as "multicultural" as they add very similar weights to ingroup and outgroup others' outcomes. On the other hand, quite a few subjects (about 36%) add not only lower but *negative* weights to the outcomes of outgroup others, displaying a very high level of ingroup bias.

In the control group in which social identity is not induced Θ^O is estimated as virtually 1. This shows that without any difference in group identities, equal weights are added to the outcomes of two random others. This finding adds confidence to the estimation method because any value significantly different from 1 would hint at a methodological artifact and cast doubt on the validity of the results. Also, the difference between the means of Θ^O in the experimental and control groups is highly significant (p 2-sided < 0.001). In the control group, the estimated variance of Θ^O is insignificant and the mean of β is estimated as 0.235. The difference in average βs in the control and

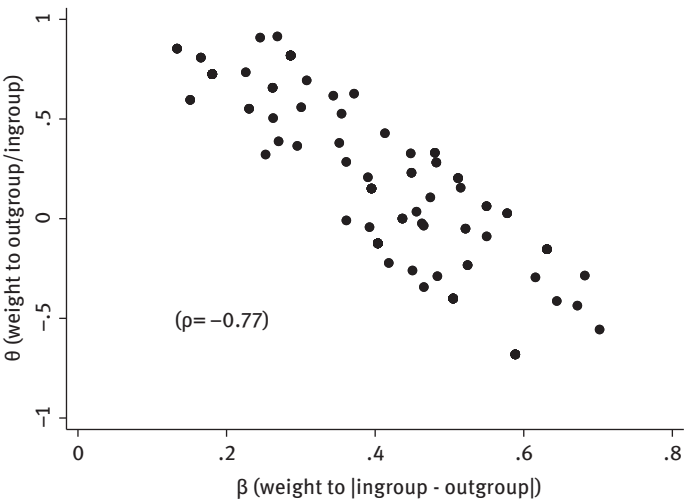


Figure 12.1: Scatter plots of utility weights in the experimental group for the two alternative interpretations of the social value orientation model.
Panel A: Θ^O and β

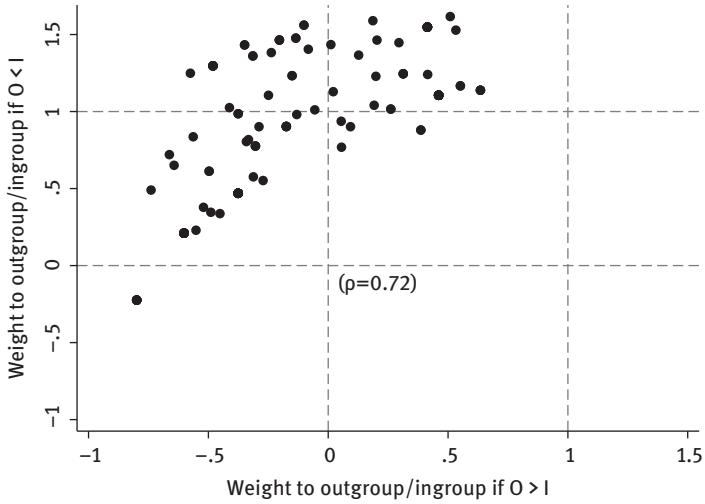


Figure 12.1 (continued)
Panel B: $[\frac{\Theta^O+\beta}{1-\beta}$ and $\frac{\Theta^O-\beta}{1+\beta}]$

experimental groups is not very large and in fact statistically marginally insignificant (p 2-sided = 0.06). Similar to the variance of Θ^O, the variance of β in the control group is statistically insignificant. Finally, the variance of the error term $(\epsilon_A - \epsilon_B)$ is virtually identical in the experimental and control groups.

A $\Theta^O = 0.2$ is a remarkably similar estimate compared to the indirect estimates of Rodrik (2017) and Kopczuk et al. (2005). The utility weight that subjects attach to the inequality between ingroup and outgroup (β), however, is an argument that Rodrik (2017) and Kopczuk et al. (2005) omit. The existence of the β term somewhat complicates the interpretation of Θ^O. This is because the actual weight one attaches to the outcome of outgroup others is affected by the inequality between ingroup and outgroup. For a more convenient interpretation, I will now use the alternative and equivalent specification of the utility function given in (12.5). In this alternative interpretation we can define two weights attached to the outcomes of outgroup others relative to that of ingroup others when (i) the outgroup is worse off than the ingroup and (ii) the outgroup is better off than the ingroup.

These weights are obtained using the β and Θ^O parameters (see equation 12.5). More precisely, when the outgroup is worse off than the ingroup (i.e., $y^I - y^O > 0$) the net weight one adds to the outcomes of outgroup others relative to ingroup others is $\frac{\Theta^O+\beta}{1-\beta}$. Substituting the estimated means of Θ^O and β gives us an estimate of $\frac{0.2+0.39}{1-0.39} = 0.97$. In other words, when outgroup others are worse off their outcomes are worth 97% of the outcomes for the ingroup, for an average subject. This means that an average subject will be somewhat indifferent to a policy that redistributes outcomes from the ingroup to the outgroup, when the outgroup is worse off. When

the outgroup is better off than the ingroup the net weight one adds to the outcomes of the outgroup relative to the ingroup is $\frac{\Theta^O - \beta}{1+\beta}$. Substituting the estimated means gives us $\frac{0.2-0.39}{1+0.39} = -0.14$. This implies that when the outgroup is in a better position than the ingroup, the average subject is spiteful: they derive negative utility from the outcomes of the outgroup other. Figure 12.1-b shows the scatter plot of these two weights. An interesting finding is the positive correlation between the two weights. The figure also shows that there is precisely one subject who adds a negative weight to the outcomes of the outgroup even when the outgroup is worse off than the ingroup. When the outgroup is better off than the ingroup, the majority of subjects are spiteful toward the outgroup.

12.5 Conclusions

In this study I bridge the social value orientation literature with the minimal group paradigm. I extend the social value orientation model to other-other allocations, particularly to the case in which the two recipients are an ingroup member and an outgroup member. Moreover, I provide a set of other-other Decomposed Games. Using these games and inducing social identity via minimal groups, I estimate the weights subjects add to the outcomes of outgroup others relative to ingroup others and to the absolute difference between the outcomes of ingroup and outgroup others. I compare these results to a control condition in which social identity is not induced. This method quantifies clearly the effect of group identity on social value orientations.

Results show that the average weight subjects add to the outcomes of outgroup others relative to ingroup others is negative when the outgroup is better off than the ingroup. In other words, the average subject is spiteful toward the outgroup when the outgroup is in an advantageous position compared with the ingroup. When the outgroup is worse off than the ingroup, the average subject weights the outcomes of ingroup and outgroup others similarly. There is also a significant variation among subjects with respect to the level of ingroup bias. While a substantial number of subjects show high levels of ingroup bias, a minority of "multicultural" subjects display little bias. This variation is an interesting finding and future research should focus on identifying subject level characteristics that explain this variation.

The method I present here is very easy to implement. The 10 other-other Decomposed Games can easily be embedded in a survey or an experimental study. Also the social identities of the two recipients in these 10 items can be adjusted depending on the researcher's interests. For example, the recipients could be from two different real social (e.g., ethnic) groups. The method I describe here gives a clear quantitative estimate of average ingroup bias. Furthermore, it captures individual differences in the level of ingroup bias. These individual-level estimates can

be outcome variables themselves. Alternatively, the researcher can use these estimates to predict an outcome variable of interest.

How much an average actor weights the outcomes of outgroups relative to ingroups affects important macro-level decisions. Barriers to international labor mobility (Rodrik 2017), foreign assistance to poor countries (Kopczuk et al. 2005), redistributive tax policies in ethnically heterogeneous contexts (Rueda 2018), intergroup trust and cooperation (Aksoy 2015; Simpson 2006), collective action in a competitive environment (Simpson and Aksoy 2017) are among the many social outcomes that are directly influenced by the extent of ingroup favoritism. The current study shows that a relatively minor minimal group identity treatment with a highly educated subject pool (the majority of the subjects were Oxford University students) is enough to create strong levels of ingroup favoritism. One could easily imagine that the extent of ingroup favoritism would be higher when real groups and less selected subject pools are considered. It is thus not surprising to see widespread support to hostile policies toward foreigners in almost all countries. Furthermore, the current study shows that ingroup favoritism is much stronger when the outgroup is better off than the ingroup. This may explain why low income groups are more likely than high income groups to favor policies that reduce the outcomes of outgroups, such as "Brexit". This is because compared with high income groups, low income groups are more likely to be worse off than the outsiders. Moreover, stressing that the outsiders are well educated and skilled may not change the opinion of the insiders. On the contrary, better off outsiders might trigger stronger ingroup favoritism from the insiders.

References

Aksoy, Ozan. 2015. "Effects of Heterogeneity and Homophily on Cooperation." *Social Psychology Quarterly* 78(4): 324–344.

Aksoy, Ozan and Jeroen Weesie. 2012. "Beliefs about the Social Orientations of Others: A Parametric Test of the Triangle, False Consensus, and Cone Hypotheses." *Journal of Experimental Social Psychology* 48(1): 45–54.

Aksoy, Ozan and Jeroen Weesie. 2013. "Hierarchical Bayesian analysis of biased beliefs and distributional social preferences." *Games* 4(1): 66–88.

Aksoy, Ozan and Jeroen Weesie. 2014. "Hierarchical Bayesian Analysis of Outcome and Process Based Social Preferences and Beliefs in Dictator Games and Sequential Prisoner's Dilemmas." *Social Science Research* 45: 98–116.

Becker, Gary. 1993. *A Treatise on the Family*. Cambridge, MA: Harvard University Press.

Fischbacher, Urs. 2007. "z-tree: Zurich Toolbox for Ready-made Economic Experiments." *Experimental Economics* 10(2): 171–178.

Greiner, Ben, 2004. "The Online recruitment system ORSEE 2.0 – A Guide for the Organization of Experiments in Economics." *Working Paper Series in Economics 10, 1–15*, mimeo, University of Cologne.

Griesinger, Donald W. and James J. Livingston. 1977. "Toward a Model of Interpersonal Orientation in Experimental Games." *Behavioral Science* 18:173–188.

Grzelak, Janusz L., Tadeusz B. Iwiński, and Jósef J. Radzicki. 1977. "Motivational Components of Utility." Pp 215–230 in *Decision Making and Change in Human Affairs*, ed. Jungermann, H., de Zeeuw, G., Dordrecht: Reidel.

Kopczuk, Wojciech, Joel Slemrod, and Shlomo Yitzhaki. 2005. "The Limitations of Decentralized World Redistribution: An Optimal Taxation Approach." *European Economic Review* 49(4): 1051–1079.

Kuhlman, Michael D., Clifford Brown, and Paul Teta. 1992. "Judgments of Cooperation and Defection in Social Dilemmas: The Moderating Role of Judge's Social Orientation." Pp: 111–132 In *Social Dilemmas, Theoretical Issues, and Research Findings*, ed. Liebrand, Wim, Dave M. Messick, and Henk A. M. Wilke, Oxford: Pergamon.

Liebrand, Wim and Charles McClintock. 1988. "The Ring Measure of Social Values: A Computerized Procedure for Assessing Individual Differences in Information Processing and Social Value Orientation." *European Journal of Personality* 2:217–230.

Macro, David and Jeroen Weesie. 2016. "Inequalities Between Others Do Matter: Evidence from Multiplayer Dictator Games." *Games* 7(2): 11.

McClintock, Charles. 1972. "Social Motivation – A Set of Propositions." *Behavioral Science* 31: 1–28.

Murphy, Ryan O., Kurt A. Ackermann, and Michael J. J. Handgraaf. 2011. "Measuring Social Value Orientation." *Judgment and Decision Making* 6(8): 771–781.

Rabe-Hesketh, Sophia, Anders Skrondal, and Andrew Pickles. 2002. "Reliable Estimation of Generalized Linear Mixed Models Using Adaptive Quadrature." *Stata Journal* 2(1): 1–21.

Rodrik, Dani. 2017. "Is Global Equality the Enemy of National Equality? *CEPR Discussion Paper No. DP11812*, url: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2908225 (acessed on 17.10.2019).

Rueda, David. 2018. "Food Comes First, then Morals: Redistribution Preferences, Parochial Altruism, and Immigration in Western Europe." *The Journal of Politics* 80(1): 225–239.

Schulz, Ulrish and Theo May. 1989. "The Recording of Social Orientations with Ranking and Pair Comparison Procedures." *European Journal of Social Psychology* 19: 41–59.

Self, Steven G., and Kung-Yee Liang. 1987. "Asymptotic Properties of Maximum Likelihood Estimators and Likelihood Ratio Tests Under Nonstandard Conditions." *Journal of the American Statistical Association* 82(398): 605–610.

Sherif, Muzafer, Harvey, O.J., B. Jack White, William R. Hood, Carolyn W. Sherif. 1961. *Intergroup Conflict and Cooperation: The Robbers Cave Experiment*. Vol. 10. University Book Exchange Norman, OK.

Simpson, Brent. 2006. "Social Identity and Cooperation in Social Dilemmas." *Rationality and Society* 18(4): 443–470.

Simpson, Brent and Ozan Aksoy. 2017. "Cumulative Advantage in Collective Action Groups: How Competition for Group Members Alters the Provision of Public Goods." *Social Science Research* 66: 1–21.

Tajfel, Henri. 1970. "Experiments in Intergroup Discrimination." *Scientific American* 223(5): 96–102.

Tajfel, Henri, Michael G. Billig, Robert P. Bundy, and Claude Flament. 1971. "Social Categorization and Intergroup Behaviour." *European Journal of Social Psychology* 1(2): 149–178.

Van Lange, Paul A. M. 1999. "The Pursuit of Joint Outcomes and Equaility of Outcomes: An Integrative Model of Social Value Orientation." *Journal of Personality and Social Psychology* 73: 337–349.

Appendix Other-other decomposed games

Table A1: 10 other-other Decomposed Games used in the study. The last three columns include percentages of subjects choosing option B in experimental and control groups and a t-statistics for the difference between the experimental and control groups, respectively (N=146 in the experimental group and N=40 in the control group).

	Option A		Option B		Data		
	(ingroup)	(outgroup)	(ingroup)	(outgroup)	% B choices		difference
game	other gets	other gets	other gets	other gets	Exp.	Control	t-value
1	9	10	11	10	.884	.925	0.747
2	10	9	10	11	.651	.900	3.124**
3	10	9	11	11	.884	.975	1.738^{+}
4	10	10	12	7	.404	.025	−4.794***
5	10	10	15	6	.616	.300	−3.671***
6	10	11	12	10	.849	.775	−1.115
7	10	14	12	8	.575	.025	−6.910***
8	11	9	10	11	.500	.925	5.155***
9	11	10	10	11	.103	.200	1.659+
10	11	11	9	10	.089	.100	0.212

***p 2-sided<0.001;**p 2-sided<0.01;$^{+}$p 2-sided<0.1

Fabian Winter and Andreas Diekmann

13 Does Money Change Everything? Priming Experiments in Situations of Strategic Interaction

Abstract: Recent work by Kathleen D. Vohs and colleagues has argued that being primed with money-related stimuli brings about a self-sufficient orientation. People reminded of money are supposedly more willing to engage in solitary vs. social activities, donate less to social causes and help others less more generally. In the spirit of Raub's triangulation efforts to understand social phenomena, we first replicate two of the individual-choice experiments reported in Vohs et al (2006). We then extend their work to strategic situations, such as the Trust Game or the Volunteers Dilemma. Our replications partially support the findings of Vohs et al. but we find no effects of money priming in strategic interaction. The results suggest that strategic, and thus social interaction may be of a different quality than individual choices. The richer context may indeed mute subtle primes such as the ones used in previous experiments.

13.1 Introduction

Rational choice theorists and analytical sociologists alike focus on three elements of explanations: Preferences, beliefs, and constraints (Hedström 2005, Gintis 2007). Both beliefs and constraints are dependent on the social structure of interaction. Thus, the social structure of interaction, the rules of 'the game', the information condition, and the payoff matrix are at the core of parsimonious explanations of social phenomena. Many situations of social interaction are strategic and the tools of game theory proved to be very useful in describing the structure precisely. In contrast to other rational choice theorists, Raub introduced game theory to sociology very early and even before the 'gold rush' of behavioral economics (e.g. Raub 1982a, 1982b, 1984). In accordance with hypothetico-deductive reasoning and with the tools of game models at hand he is able to derive precise predictions which can be corroborated by experimental methods or survey data. Raub's pioneering work on social

Note: The title is adapted from Goetzmann (2016).

Fabian Winter, MPRG Mechanisms of Normative Change, Max Planck Institute for Research on Collective Goods, Bonn, Germany
Andreas Diekmann, Universität Leipzig and ETH Zurich, Zürich, Switzerland

https://doi.org/10.1515/9783110647495-013

dilemmas, on networks and social exchange, on repeated games in a sociological context, on trust and commitment and many other topics demonstrated the fruitful approach of the "Utrecht school" (e.g. Raub 2004; Raub and Weesie 1990; Raub and Buskens 2011). On the other hand, psychological game theory and framing models may extend the parsimonious models. By this approach, more or less salient motives are elicited by priming techniques. Also, the perception of the situation can be manipulated by presenting a certain type of "frame" (e.g. Esser 2010; Lindenberg 2006 for framing theories in sociology). A typical example is the "Wall Street Game experiment" (Liberman et al. 2004). A switch from a "Community" to a "Wall Street" frame strongly reduced the probands' inclination to contribute to the common goal in a Prisoner's Dilemma.

In a similar vein, Vohs et al. (2006, 2008) suggested a priming theory of money. They argue that a money prime (e.g. reading a text, a list of words referring to monetary transactions or observing images reminding of money, etc.) elicits cognitive processes and activities that may be described as more "self-sufficient". The self-sufficiency hypothesis does not necessarily imply selfish behavior. The authors predict that "when reminded of money, people would want to be free from dependency and would also prefer that others not depend on them" (Vohs et al. 2006). On the other hand, money priming may enhance performance and an equity, exchange-based orientation (Vohs et al. 2008).

The effects of money priming are demonstrated by nine experiments (Vohs et al. 2006; further experiments are reported in Vohs et al. 2008). All experiments refer to non-strategic social interactions, i.e. the outcome of a decision is not dependent on other actors' strategies. The question arises if money priming may also have self-sufficiency effects in situations of strategic interactions, as modelled by game theory. These are clearly predicted by the hypothesis because money-primed subjects should focus more on oneself and less on others. In the following we will try to replicate two experiments conducted by Vohs et al. (2006). In a second step we will extend this research by investigating money priming in four games of strategic interaction: The Ultimatum Game (UG), the Trust Game (TG), the Prisoner's Dilemma (PD) and the Volunteer's Dilemma (VOD).

13.2 The experiments by Vohs et al

In this section, we want to briefly summarize the relevant experiments and results presented in the work by Vohs et al. (2006). In total, the authors present nine experiments. We will later repeat two of them, although we follow different experimental protocols than the original authors. Our study is thus not a direct one-to-one replication but an independent test of their theory.

Here is the description of the design of the questionnaire study taken from the original publication:

> In Experiment 8, we tested whether money-primed participants would place a premium on being alone even when choosing leisure activities that could be enjoyed with friends and family. Participants were randomly assigned to one of three priming conditions. Participants first sat at a desk, which faced one of three posters, to complete filler questionnaires. In the money condition, the desk faced a poster showing a photograph of various denominations of currency (fig. S3). In two control conditions, the desk faced a poster showing either a seascape or a flower garden (figs. S4 and S5). Subsequently, participants were presented with a nine-item questionnaire that asked them to choose between two activities. Within each item, one option was an experience that only one person could enjoy and the other option was for two people or more (e.g., an in-home catered dinner for four versus four personal cooking lessons). (Vohs et al. 2006)

The authors find that participants in the money-prime condition are more likely to choose solitary activities than those in the two control conditions.

The second replication in this chapter zeros in on the relation between money and altruism. Here is the description of the original experiment:

> Experiment 6 tested for the psychological effects of money by operationalizing helpfulness as monetary donations. Upon arrival to the laboratory, participants were given $2 in quarters in exchange for their participation. The quarters were said to have been used in an experiment that was now complete; in actuality, giving participants quarters ensured that they had money to donate. Participants were randomly assigned to one of two conditions, in which they descrambled phrases that primed money or neutral concepts. Then participants completed some filler questionnaires, after which the experimenter told them that the experiment was finished and gave them a false debriefing. [. . .] As the experimenter exited the room, she mentioned that the lab was taking donations for the University Student Fund and that there was a box by the door if the participant wished to donate. Amount of money donated was the measure of helping. (Vohs et al. 2006)

As predicted by their theory, the authors find that participants primed with money donate significantly less than those in the neutral frame.

Our replications deviate in several aspects from the original studies, which we further spell out in the next section. Importantly, we use the screen-saver priming from the donation experiment also in the questionnaire study instead of the posters used by Vohs et al. (2006).

13.3 Experimental Setup

Our experiments were conducted as computerized experiments with the zTree package (Fischbacher 2007) in the experimental lab of the ETH Zürich. Participants were students from the University of Zürich and the ETH Zürich who were invited via ORSEE, an internet-based recruitment system (Greiner, 2015) and had never participated in an economic experiment before. As it is standard in the experimental research

on social dilemmas and strategic interaction, all experiments were incentivized, however in a non-standard way. Instead of playing for money, our participants played for time in the lab (see for instance Berger et al. 2012 for a similar approach). Participants received a flat payment of 33 SFr. and played over the amount of simple mathematical calculations everybody had to perform. The more of these calculations someone was assigned to as the outcome of the games, the longer it took him or her to perform the task. Due to this design choice, we were able to induce an incentivized strategic situation without relating it to money, which could threaten the treatment effect.

All experiments involved the same priming mechanism adopted from Vohs et al. (2006), which worked as follows. First, participants did 100 calculations to familiarize with the "currency". Thereafter, they were asked to fill out a questionnaire including the experimental questions for experiment 1 and to read the instructions for either experiment 2 or experiment 3, which took about 7–15 minutes. During that time, a screensaver appeared, which either showed fish (neutral prime) or money bills (money prime) floating through water (see Supporting Online Material of Vohs et al. 2006). Every participant took part in experiment 1 and either in experiment 2 *or* experiment 3.

13.4 Findings

13.4.1 Experiment 1: Replication of the Vohs et al. questionnaire experiment

Our first experiment was designed to test the robustness of experiment 8 reported in of Vohs et al. (2006) and to confirm that our manipulation worked. We included the same experimental items in our questionnaire, asking participants to choose between nine social and solitary activities (e.g., an in-home catered dinner for four versus four personal cooking lessons). Confirming the results of Vohs et al. (2006), we find that participants primed on money ($N = 66$) choose significantly more solitude activities as compared to those with a non-money prime ($N = 66$) [$\mu_{money} = 2.97$ ($\sigma = 1.15$), $\mu_{neutral} = 2.53$($\sigma = 1.26$), $t(130) = 2.10$, $p = 0.038$, Cohen's $d = 0.371$]. Note, however, that the effect size is much smaller than in the Vohs et al. study and at an overall lower level.[1] We are thus confident, that our priming worked.

1 Vohs et al.: $\mu_{money} = 4.00$ ($\sigma = 1.20$), $\mu_{neutral1}/\mu_{neutral2} = 2.82/3.10$ ($\sigma_{neutral1} = 1.00/\sigma_{neutral2} = 1.80$), Cohen's d money vs. neutral 1 =0.59, Cohen's d money vs. neutral 2 = 1.06.

13.4.2 Experiment 2: Does activating the concept of money affect fairness and trust?

Experiments 2 directly followed experiment 1 in two sessions with 30 participants each ($N = 60$). According to experiment 6 in Vohs et al. (2006), priming participants on money reduces their willingness to donate money to charitable organizations. With our second experiments, we aim to test the robustness of this finding in a more controlled environment. To get a robust measure of altruistic and fair behavior, we implemented the Ultimatum Game (UG, Güth et al. 1982), the Trust Game (TG, Dasgupta, 1988) and the Dictator Game (DG, Forsythe et al. 1994), played once by all participants in this order. If the results of Vohs et al. (2006) hold, we would expect less egalitarian behavior across games if participants are primed on money.

In UG, two players, *Proposer* and *Responder*, bargain over a divisible good (in our case 100 calculations). The Proposer makes an offer, stating how many of the 100 calculations the Respond41er has to solve. At the same time, the Responder states how many calculations he or she will solve *at most*, without knowing the Proposers offer. If the Proposer's offer is at least as low as the Responder's threshold, the Responder will solve the offered calculations, and the Proposer will solve the remaining ones. This procedure is known as the strategy method (Selten 1967) and has been argued to elicit normative behavior (Rauhut and Winter 2010). If the offer is above the threshold, both players have to solve the full 100 calculations.[2] Under the canonical assumptions of game theory, the subgame perfect equilibrium of the game is such that the responder will accept any positive amount, and by backward induction the proposer offers exactly that. A higher threshold of the responder is often associated with a concern for equitable outcomes (cf. Fehr and Schmidt 1999; Winter et al. 2012). Money-induced self sufficiency should thus lead to lower thresholds, as the comparison with the Proposer is less important. Conversely, a self-sufficient Proposer should make lower offers, as fairness as well as the Responder's fairness concerns are less important. In both cases, however, we cannot reject the null hypothesis of different offers and thresholds, respectively [Offers: $\mu_{money}) = 44.6$ ($\sigma = 16.8$), $\mu_{neutral}) = 45.7$ ($\sigma = 11.0$), $t(130) = 0.29$, $p = 0.77$, Cohen's $d = .08$, Thresholds: $\mu_{money}) = 51.7$ ($\sigma = 17.1$), $\mu_{neutral}) = 54.1$ ($\sigma = 12.4$), $t(130) = 0.62$, $p = 0.53$, Cohen's $d = .16$].

In TG, there are again the two roles of Proposer and Responder. First, the Proposer decides between the two options *C* (for cooperate) and *D* (for defection). If the Proposer chooses *D*, both players have to solve 200 calculations. If the Proposer chooses *C*, the number of calculations depends on the Responder's decision: If the Responder chooses *C*, both players have to solve 150 calculations, which is less

2 In this and the following two games, we used the *strategy vector method*, that is, we asked the participants to decide for both roles, Responder and Proposer, and only later assign them randomly to one of the roles. See Rauhut and Winter (2010) for a discussion of this method.

than the 200 they would have to solve if the Proposer chooses *D*. If the Responder chooses *D*, however, he or she would have to solve no calculations at all, while the Proposer has to solve 300. Thus, if the Proposer trusts in the Responder's fairness, he would choose *C*, but *D* otherwise. Again, via backward induction, we can show that the (subgame perfect) Nash equilibrium of the game leads the proposer to choose *D*, yielding (200, 200) because the responder would choose *D* herself, yielding (300, 0) which is worse than the 200 calculations the proposer would receive from opting out early.

Self sufficiency should lower the willingness to rely on the friendly behavior of others and should thus lower both player's probability of choosing *C*. Again, we are not able to reject the null hypothesis for either of the decisions [Proposer: $\chi^2(1, n = 30) = 0.07$, $p = 0.60$, Responder: $\chi^2(1, n = 30) = 0.27$, $p = 0.79$].

The DG mirrors the UG, however without the Responder's power to veto an unacceptable offer. Thus the Proposer's task is simply to decide how to allocate 100 calculations between himself/herself and the Responder. The DG can therefore be considered as a replication of experiment 6 in Vohs et al. (2006). But also in this case, we do not find significant differences between the two experimental conditions [Offers: $\mu_{money} = 67.2(\sigma = 24.7)$, $\mu_{neutral} = 65.6(\sigma = 24.5)$, $t(130) = 0.25$, $p = 0.80$, Cohen's $d = .07$].

In sum, we cannot find evidence that participants primed on money behave differently from those confronted with a neutral prime in strategic interactions.

13.4.3 Experiment 3: Does activating the concept of money affect cooperation?

Experiments 3 directly followed experiment 1 in four other sessions with 18–20 participants each ($N = 74$). As in experiment 2, participants were exposed to the priming and answered a set of control questions. The experimental task consisted of two other well-know experimental paradigms, the Prisoner's Dilemma (cf. Rapoport and Chammah 1965), and the Volunteer's Dilemma (Diekmann 1985) displayed in Table 13.1. Both games were played once in the above order by all the participants without giving feedback about the outcomes until the very end of the experiment.

Table 13.1: Prisoner's Dilemma (left) and Volunteer's Dilemma (right). The values in the cells denote the number of calculations to be performed associated with the choices.

Player 1 \ Player 2	Cooperation	Defection
Cooperation	200, 200	200, 0
Defection	0, 200	300, 300

Player 1 \ Player 2	Cooperation	Defection
Cooperation	100, 100	300, 0
Defection	0, 300	200, 200

In the Prisoner's Dilemma (PD), defection is always better on the individual level than cooperation (denoted by C), but worse on the collective level, which creates the social dilemma. Self sufficiency should thus lead to less cooperation, as people tend to focus on individual rather than collective aims. Our empirical results, however, show an unexpected effect in the opposite direction: Priming participants on money *increases* the likelihood of cooperation ($\chi^2(1, n = 74) = 4.34, p = 0.04$).

On the contrary, cooperation should be more pronounced in the Volunteer's Dilemma (VD). The game as played by our participants has three equilibria: two in pure strategies where either of the two cooperates while the respective other abstains, and a mixed strategy equilibrium where both players cooperate with the same probability $0 < p < 1$. From the decision maker's perspective, the most important thing is that *someone* (including the decision maker) volunteers to choose C, but its better if someone else does it. In this case, money-induced self sufficiency would lower the willingness to rely on someone else, and consequently increase the likelihood of choosing the cooperative option. Our empirical results, however, do not allow us to reject the null-hypothesis of no differences between treatments: Participants primed on money are as cooperative as participants exposed to the neutral prime ($\chi^2(1, n = 74) = 0.05, p = 0.82$). In sum, we cannot confirm the hypothesized effects of money-priming on cooperation levels in incentivized social dilemmas.

13.5 Conclusions

We replicated the effect of a money screensaver on the choice of solitary activities (experiment 8 in Vohs et al. 2006). The effect size found in our replication was lower than in the original experiment. The diminished size of the effect is no surprise. Smaller effect sizes found in replications are very well in accordance with "the winner's curse" hypothesis (Young et al. 2008) as well as with results from the recent wave of replications in psychology (Klein et al. 2014; Open Science Collaboration 2012, 2015). As in commercial auctions, where of course the largest bid will win, large experimental effects found in scientific research have an increased chance of publication. Due to publication bias replication studies often report smaller effect sizes than the original study (if there is an effect at all). Moreover, we did not find a significant difference in fairness behavior with the Dictator Game. The replication of experiment 6 of Vohs et al. (2006) failed to provide evidence for the self-sufficiency hypothesis. Our results also mirror recent replication attempts of the Vohs et al. study by Klein et al. (2014), who also find no effects of the prime across many labs.

Our key questions target possible effects of money priming on the behavior in situations of strategic interaction. Here, our answer is clearly negative. In the Ultimatum Game neither the proposer's nor the responder's behavior was significantly influenced by the money screen saver. Also, there was no significant effect

in the Trust Game or in the Volunteer's Dilemma. For the Prisoner's Dilemma we even found a significant effect in the opposite direction of what was predicted. Here, money-primed subjects showed significantly more cooperation than subjects in the control group.

At least in situations of strategic interaction, the evidence for money-priming effects on actor's decisions is weak or non-existent. This is not to say that priming or framing has no impact on strategic interactions in general. For example, the Wall Street Game experiment provides evidence to the contrary. Framing the situation of a Prisoner's Dilemma had strongly affected the level of cooperation (Liberman et al. 2004). However, we do not know under which conditions which type of priming or framing really has an impact on the various situations of strategic interaction.[3]

In contrast, game theory is much more systematic and elaborated. Given the structure of strategic interaction, game theory is a powerful tool to derive precise predictions on actors' decisions. The work of Raub and collaborators clearly demonstrates this. Enriching game models by social context, framing, and psychological hypotheses is an important step to enhance the explanatory power of a theory of social interaction. However, we are still far away from a general, unified theory and doubts remain that the social sciences will ever attain this goal.

References

Berger, Roger, Heiko Rauhut, Sandra Prade and Dirk Helbing (2012). Bargaining over Waiting Time in Ultimatum Game Experiments. *Social Science Research*, 41(2), 372–379.

Dasgupta, Partha (1988). Trust as a commodity. In Gambetta, Diego, editor, *Trust*, chapter 4, pages 49–72. Oxford: Blackwell Publishers.

Diekmann, Andreas (1985). Volunteer's dilemma. *The Journal of Conflict Resolution*, 29(4):605–610.

Dreber, Anna, Tore Ellingsen, Magnus Johannesson, David G. Rand (2013). Do People Care About Context? Framing Effects in Dictator Games. *Experimental Economics*, *16*(3), 349–371.

Esser, Hartmut, (2010). Das Modell der Frame-Selektion. Eine allgemeine Handlungstheorie für die Sozialwissenschaften? In G. Albert, S. Sigmund, Eds., Soziologische Theorie kontrovers. *Kölner Zeitschrift für Soziologie und Sozialpsychologie* 50:45–62.

Fehr, Ernst and Klaus M. Schmidt (1999). A Theory of Fairness, Competition and Cooperation. *Quarterly Journal of Economics*, 114(3):817–868.

Fischbacher, Urs (2007). z-Tree: Zurich Toolbox for Ready-Made Economic Experiments. *Experimental Economics*, 10(2):171–178.

Forsythe, Richard, Joel Horowitz, N.E. Savin, and Martin Sefton (1994). Fairness in Simple Bargaining Experiments. *Games and Economic Behavior* 6(3):347–369.

3 An interesting question is whether the strong framing effect of the Wall Street Game might diminish or completely vanish when there are increasing stakes. Is "structure", i.e. the structure of the game more important than priming or framing, particularly when stakes are high? Is a possible effect of framing due to a change of preferences or the beliefs or both? (Dreber et al. 2013).

Gintis, Herbert (2007). A Framework for the Unification of the Behavioral Sciences. *Behavioral and Brain Sciences* 30:1–61.

Goetzmann, William N. (2016). *Money Changes Everything. How Finance Made Civilization Possible*. Princeton & Oxford: Princeton University Press.

Greiner, B. (2015). Subject pool recruitment procedures: organizing experiments with ORSEE. *Journal of the Economic Science Association*, *1*(1), 114–125.

Güth, Werner, Rolf Schmittberger, and Bernd Schwarz (1982). An Experimental Analysis of Ultimatum Bargaining. *Journal of Economic Behavior & Organization* 3(4):367–388.

Hedström, Peter (2005). *Dissecting the Social. On the Principles of Analytical Sociology*. Cambridge: Cambridge University Press.

Klein, Richard A., Kate A. Ratliff, Michelangelo Vianello, Reginald B. Adams, Štěpán Bahník, Michael J. Bernstein, . . . and Brian A. Nosek, (2014). Investigating variation in replicability. *Social Psychology* 45(3):142–152.

Liberman, Varda, Steven M. Samuels, and Lee Ross (2004). The Name of the Game. Predictive Power of Reputations Versus Situational Labels in Determining Prisoner's Dilemma Game Moves. *Personal and Social Psychology Bulletin* 30:1175–1185.

Lindenberg, Siegwart (2006). Prosocial Behavior, Solidarity, and Framing Processes. In Fetchenhauer, D., Flache, A., Buunk, B., Lindenberg, S., Eds., *Solidarity and Prosocial Behavior. An Integration of Sociological and Psychological Perspectives*, pages 23–44. New York: Springer.

Open Science Collaboration. (2012). An open, large-scale, collaborative effort to estimate the reproducibility of psychological science. *Perspectives on Psychological Science* 7(6):657–660.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349* (6251), aac4716.

Rapoport, Anatol and Albert Chammah (1965). *Prisoner's Dilemma*. The University of Michigan Press, Ann Arbor.

Raub, Werner (1982a). *Theoretical Models and Empirical Analyses. Contributions to the Explanation of Individual Actions and Collective Phenomena*. Utrecht: Explanatory Sociology Publications.

Raub, Werner (1982b). The Structural-Individualist Approach: Towards an Explanatory Sociology, *in* Raub, Werner, editor, *Theoretical Models and Empirical Analyses* (Utrecht, E. S. Publications, pp. 3–40).

Raub, Werner (1984). *Rationale Akteure, institutionelle Regelungen und Interdependenzen. Untersuchungen zu einer erklärenden Soziologie auf strukturell-individualistischer Grundlage*. Frankfurt am Main: Peter Lang.

Raub, Werner and Jeroen Weesie (1990). Reputation and Efficiency in Social Interactions. *American Journal of Sociology* 96:626–654.

Raub, Werner (2004). Hostage Posting as a Mechanism of Trust. Binding, Compensation, and Signaling. *Rationality and Society* 16:319–365.

Raub, Werner and Vincent Buskens (2011). Micro-Macro Links and Microfoundations in Sociology. *Journal of Mathematical Sociology* 35:1–25.

Rauhut, Heiko and Fabian Winter (2010). A sociological perspective on measuring social norms by means of strategy method experiments. *Social Science Research*, 39(6):1181–1194.

Selten, Reinhard (1967). Die Strategiemethode zur Erforschung des eingeschränkt rationalen Verhaltens im Rahmen eines Oligopolexperiments. In Heinz Sauermann, Ed., *Beiträge zur experimentellen Wirtschaftsforschung*, Vol. I, pages 136–168. Tübingen: J.C.B. Mohr – Siebeck.

Vohs, Kathleen D., Nicole L. Mead, and Miranda R. Goode (2006). The psychological consequences of money. *Science*, 314:1154–1156.

Vohs, Kathleen D., Nicole L. Mead, and Miranda R. Goode (2008). Merely Activating the Concept of Money Changes Personal and Interpersonal Behavior. Current Directions in *Psychological Science* 17:208–212.

Winter, Fabian, Heiko Rauhut, and Dirk Helbing (2012). "How norms can generate conflict: An experiment on the failure of cooperative micro-motives on the macro-level". *Social Forces*, 90: 919–948.

Young, Njoo S., John P. A. Ioannidis, Omar Al-Ubaydli (2008). Why Current Publication Practices May Distort Science. *PLoS Medicine* 5:1418–1422.

Martin Abraham, Kerstin Lorek and Bernhard Prosch

14 Social Norms and Commitments in Cooperatives – Experimental Evidence

Abstract: Cooperatives, which are characterized by pooling of jointly owned and controlled resources in an enterprise by individual actors, are popular and widespread in modern societies. However, since each actor has an incentive to withhold resources individually while benefiting from the common pool, opportunistic behavior may result. One possibility to overcome this dilemma situation are internalized, normative beliefs which foster cooperative behavior. By experimentally modeling dilemma situations, we examine whether normative values work as behavioral reference points for members of cooperatives and whether this enhances cooperation. Our results from two lab experiments demonstrate that a cooperative framework, which we use as an indicator for normative beliefs, produces significantly higher cooperation rates in social dilemma situations. Furthermore, we see that an institution framed as a cooperative is chosen by a substantial share of persons, even if this institution produces inefficient results. Consequently, we conclude that general norms contribute to the cooperative effect of cooperatives.

14.1 Introduction

Cooperatives are popular around the world. Rooted in forms of collective action already in medieval times (De Moor 2008), it became also an important type of organization in capitalistic systems. Beginning with the Rochdale Society of Equitable Pioneers in 1844 in England, the idea of cooperatives as an organizational form of self-help for bundling resources and power spread around the world and resulted in a multitude of cooperatives. The International Co-operative Alliance defines a cooperative as "[. . .] an autonomous association of persons united voluntarily to meet their common economic, social, and cultural needs and aspirations through a

Notes: We gratefully acknowledge financial support from the Ludwig-Erhard-Forschungsgesellschaft (fund AO 7202119, principal investigator: Bernhard Prosch).

This paper is based on a research project proposed and started by Bernhard Prosch. Since he died far too early in 2015, the project was finished by the co-authors. The commitment experiment in this paper was completely carried out by Bernhard Prosch, and the framing experiment was carried out by the co-authors. We dedicate this paper to Bernhard, who was a passionate sociologist and experimenter. In addition we want to thank Miriam Rudel for support during the experiments.

Martin Abraham, Kerstin Lorek, Bernhard Prosch, Friedrich-Alexander University, Erlangen-Nürnberg

https://doi.org/10.1515/9783110647495-014

jointly owned and democratically-controlled enterprise" (ICA 2018). All cooperatives rely on an internationally shared set of values rather than profit orientation. The shared values comprise self-help, self-responsibility, democracy, equality, equity, and solidarity.

Cooperatives can be found in various economic sectors (Higl 2008: 9f). Some examples are credit unions, housing cooperatives, energy cooperatives or agricultural cooperatives. In Europe, approximately 176,000 cooperatives exist, with 141 million members and 4 million employees. Since 2009, the number of cooperative enterprises has increased by approximately 12%. Italy, Turkey, France, and Spain comprise the countries with the largest numbers of cooperatives. The largest number of cooperative members can be found in France, Germany, the Netherlands, the UK, and Italy. Most of the cooperatives in Europe are in the industry and services sector, followed by agriculture and housing. The largest sector by membership is, however, banking, followed by consumer and insurance.

In Germany, there are almost 7,500 cooperatives that have approximately 22 million members and an annual turnover of 195 € billion. In 2015, 126 new cooperatives were founded. The largest sector is agriculture, followed by the housing, industry, services, and social sectors as well as banking. In the Netherlands, approximately the same number of people (20 million) are members of one of the 70 cooperatives, and agriculture is also the largest sector (all numbers come from Cocolina 2016).

In general, people find and join cooperatives to overcome problematic situations by pooling their resources. Theoretically, in these situations, a "social dilemma" arises, which is also often called a public good dilemma. Actors face cooperation problems: since each actor has an incentive to withhold resources individually while using the common pool of resources, group and individual interests diverge, and opportunistic behavior may result. One of the classic examples is the tragedy of commons (Hardin 1968), where, for example, farmers overexploit land owned by the community because cattle are sent to graze there by all farmers. Institutions offer a solution for social dilemma situations by establishing rules to enforce cooperation.

In this paper, we see cooperatives as a specific kind of institution that bundles resources and provides incentives for cooperative behavior in dilemma situations. In our opinion, there are basically two types of mechanisms that may lead to cooperative behavior in cooperatives. First, cooperatives may be based on general normative beliefs, which promote cooperation in cooperative institutions. Second, sanctions of misbehaving members are the classical mechanisms to ensure cooperative behavior (see, for example, Camerer 2003; Raub 2004; Raub, Buskens, and Corten 2015).

The last mechanism has been the main subject of research on cooperation, and selective incentives such as sanctions, hostages or rewards can enhance cooperation. However, often enough, the assignment of this mechanism is difficult due to information problems or the possibility to guarantee a sufficiently high incentive. Especially, when we look at the literature on cooperatives, we see that norms are an important part of this type of organization (for example, Spear 2000: 520). However, empirical evidence on the effectiveness of such mechanisms is rare.

Consequently, the aim of our paper is to analyze the role of normative beliefs in the functioning of cooperatives. By experimentally modeling dilemma situations, we examine whether normative values work as behavioral reference points for members of cooperatives and whether this enhances cooperation. Our results demonstrate that a cooperative framework, which we use as an indicator for normative beliefs, produces significantly higher cooperation rates in social dilemma situations. Furthermore, we see that an institution framed as a cooperative is chosen by a substantial share of persons, even if this institution produces inefficient results. Consequently, we conclude that general norms contribute to the cooperative effect of cooperatives.

In the following sections, we take a closer look at the concept of cooperatives and their principles (2). We then argue that cooperative principles work as an informational framework, helping build expectations about others' behavior and therefore affect their own behavior (3). Furthermore, institutional rules of cooperatives work as (costly) commitment devices that enhance cooperative behavior (3). We describe our experiments and their results (4) and conclude with a discussion of the implications of our findings (5).

14.2 Basic principles of cooperation in cooperatives

Regardless of cooperatives' specific interest, all cooperatives share one basic idea: the collective solution of problems via cooperation and a set of shared values. Although there are examples of cooperatives without primary economic goals (for example, cooperatives that organize living opportunities for senior citizens), most cooperatives have economic aims. The basic principles that any cooperative is based on also form a distinctive difference from conventional profit-maximizing enterprises. Although within more economically oriented cooperatives, members expect to make a profit, and the maximum support for its members and cooperatives' causes remain the main goal. The generated surplus is distributed evenly among the members or reinvested in the cooperative (Draheim 1955: 95f). As a result, each individual member of the cooperative benefits from collectively produced resources and services. The cooperative pools resources and makes it possible to reach goals that would not be possible with only the individual's capacity. The core principles of self-help, self-administration, self-responsibility, and the identity principle underline this notion.

According to the *principle of self-help*, members of cooperatives share a common interest that they follow based on voluntary and open membership by cooperating with each other. The underlying goal is the promotion of their economic and social situation without external help and as a solidary collective. The principle of self-help therefore implicitly formulates the cooperative values of *subsidiarity and solidarity* (Harbrecht 2000).

The *principle of self-administration* states that cooperatives are the property of their members. The management of cooperatives therefore lies within the hands of their members. Executive directors and the supervisory board are therefore elected by all members at general meetings. Decisions are based on the one member, one vote principle, which awards every cooperative member with only one vote, regardless of the member's investment.

The *principle of self-responsibility* states that the cooperative and its members take over sole liability against its creditors (Zerche, Schmale, and Blome-Drees 1998: 9ff). Bonus (1994: 64ff.) formulates additional values: equality, equal treatment, trust and reliability. All of this goes along with the *identity principle*. According to this principle, each member of the cooperation fulfills a double role as owner and employee, client, and supplier at the same time.

These key principles, which are identified in the literature as the core principles of cooperatives, can be seen as specific norms that guide the individual behavior of cooperative members. We believe that people know in advance that a membership requires compliance with these norms, and members will tend to cooperate due to internalized norms and selective membership. We will describe this idea in more detail in the next section.

14.3 Cooperatives and the solution of social dilemma situations

There are basically two mechanisms that may explain the higher cooperation rates within cooperatives. First, cooperatives may have a normative basis, which is based on internalized beliefs about appropriate behavior towards other members of a cooperative (Spear 2000: 520). Second, the sanctioning of misbehaving members and incentives for compliant members are the classical mechanisms for ensuring cooperative behavior (see, for example, Camerer 2003; Raub 2004; Raub, Buskens, and Corten 2015). Moreover, those incentives are safeguards against defective free riders seeking to exploit normative-driven cooperators. Often, the sanctions and incentives result from network embeddedness, for example, by awarding status or excluding defecting members from other private activities (Raub and Weesie 1990; Fehl, van der Post, and Semmann 2011; Raub, Buskens, and Frey 2013). Although this network embeddedness will probably have in most cooperatives an important effect (Ole Borgen 2001), there are cooperatives that do not rely heavily on face-to-face contact among the members, such as members of a banking cooperative. In this paper, we question whether cooperatives can increase cooperation, even without close-knit networks among members. Consequently, we concentrate on the effect of normative beliefs tied to cooperatives.

14.3.1 Cooperative values as informational cues for cooperative behavior

Dilemma situations often arise because in the short run, opportunistic behavior is more profitable for individuals than cooperation. In the long run, however, the group – and therefore an individual – benefits from cooperation (see, for example, Raub and Voss 1986; North 1990). Institutions and norms are the core mechanisms that help to overcome opportunistic behavior and support cooperation among individuals (Coleman 1990: Ch. 10; Hechter, Opp, and Wippler 1990; North 1990; Williamson 1985).

In this paper, we analyze cooperatives as a special kind of institution. People associate and organize to solve a common problem such as making an investment that would be too large for a single actor to perform. The core values and principles of cooperatives describe social norms that expect a cooperative's member to invest and to cooperate to reach a commonly shared goal. However, even within cooperatives, incentives for opportunistic behavior exist. Each member's contribution within the cooperative can only be observed incompletely by the other members and management. This leads to asymmetrical information problems. For an individual actor, incentives exist to maximize his or her own profit from the collective by reducing his or her own contribution (Olson 1965; Akerlof 1970: 490ff).

The result would be collectively inefficient, and cooperatives therefore would not be a stable organizational form. However, as we observe a decades-long cooperative tradition, there must exist internal structures or mechanisms that create incentives for cooperative behavior and promote norm-abiding behavior among its members.

The role as a member of a cooperative is accompanied by formal and informal expectations about the behavior fueled by the cooperative principles. This in turn influences other members' behavior. Within cooperatives, everyone knows what behavior can be expected and what behavior would not be accepted. This argument is supported by Draheim (1955: 43ff), who sees the "ghost of the cooperatives" promoting loyalty among members and reducing opportunism (see also Spear 2000: 520).

We argue that the use of the term "cooperative" in an experiment can trigger these (internalized) normative beliefs. We follow here Dufwenberg, Gächter, and Hennig-Schmidt (2011: 472), who define framing effects as "a two-part process where (i) frames move beliefs, and (ii) beliefs shape motivation and choice." They find that the use of the term "community" can lead to higher cooperation rates in a Prisoner's Dilemma. An additional explanation is provided in the "Wall Street versus Community Game" by Liberman, Samuels, and Ross (2004) and Ellingsen et al. (2012), who similarly find that cooperative behavior is more likely when the game is called the "Community Game" than when it is called the "Wall Street Game". However, they conclude that social frameworks can be coordination devices that influence people's beliefs rather than their preferences. In this sense, cooperation would increase because actors expect others to be more cooperative in an appropriate framework.

We do not distinguish between beliefs and preferences in our paper, and both arguments lead to the hypothesis that participants should behave more cooperatively if they are given the possibility to act under a "cooperative" framework.

Moreover, normative beliefs may affect not only cooperative behavior but also the choice to become a member of an institution. If these norms are sufficiently strong, they may even override a primary incentive structure. We will use this possible effect to reveal the normative motives when deciding to join a cooperative. Therefore, we will look at a situation in which the cooperative solution produced by a cooperative is inefficient and observe to what extent people nevertheless join the institution.

14.4 Experiments

14.4.1 Framing experiment

The first experiment examines the effect of membership in cooperatives and salient knowledge of the cooperative principles on cooperation. The experiment took place at the LERN (Laboratory for Experimental Research Nuremberg at the Department of Economics of the University of Erlangen-Nürnberg), with a total of 84 student participants. Programming and execution were performed with z-Tree (Fischbacher 2007), and recruiting was performed with ORSEE (Greiner 2004).

The participants are, on average, 24 years old and in their third year of study. Approximately 60% of the participants had knowledge about the concept of cooperatives beforehand. Approximately the same share of participants could imagine joining a cooperative. This positive attitude towards cooperatives is further supported by two-thirds of the participants thinking of cooperatives as a good idea.

The experiment was a basic one-shot public good game (first introduced by Marwell and Ames 1979) for the control group (CG), with three treatments, conducted in a total of three sessions. For each session, participants were randomly invited using the participant pool of the LERN. In the control treatment (CT), the control group, each participant receives a starting budget. Divided into groups of four, each participant has to contribute to a group project. The sum of the individual contributions within the group is multiplied by 1.5, and the resulting amount is divided equally among the group members. Economic theory based on an egoistic actor would predict a cooperation of 0 from participants. Empirically, however, various studies show cooperation rates higher than 0 (for example, Fehr and Fischbacher 2004; Fehr and Gächter 2000).

Treatment 1 (T1) includes a framing treatment. In addition to the instructions of the Public Good Game, participants receive information about the concept of cooperatives and the cooperative value system (the text of this instruction can be found in the Appendix of this chapter). Participants are again divided into groups

of four. However, the groups are called cooperatives throughout the instructions and experiment. With this treatment, we examine the effect of the normative beliefs triggered by being a member in a cooperative.

Treatment 2 (T2) adds a mandatory contribution of 10% of the starting budget. This mandatory contribution is added to the group contribution and again equally divided between the group members. In all other aspects, this treatment resembles T1. This treatment considers the investments that come along with the membership in a cooperative. As one of the goals of cooperatives is to collect enough resources to reach the shared (economic) goal, we test whether a mandatory contribution leads to the crowding out effects of voluntary cooperation.

The descriptive analysis of this first experiment (see Figure 14.1) shows significantly higher cooperation for T1 (66.8%) than for the control group (47.1%). This indicates that, according to our expectation, that the framing effect of being in a cooperative instead of a free group and consequently the triggering of cooperative values increase cooperation. The cooperation rate in T2, 62.3%, only slightly and not significantly differs from the rate in T1. The mandatory contribution seems to have no negative effects on participants' intrinsic motivation to contribute.

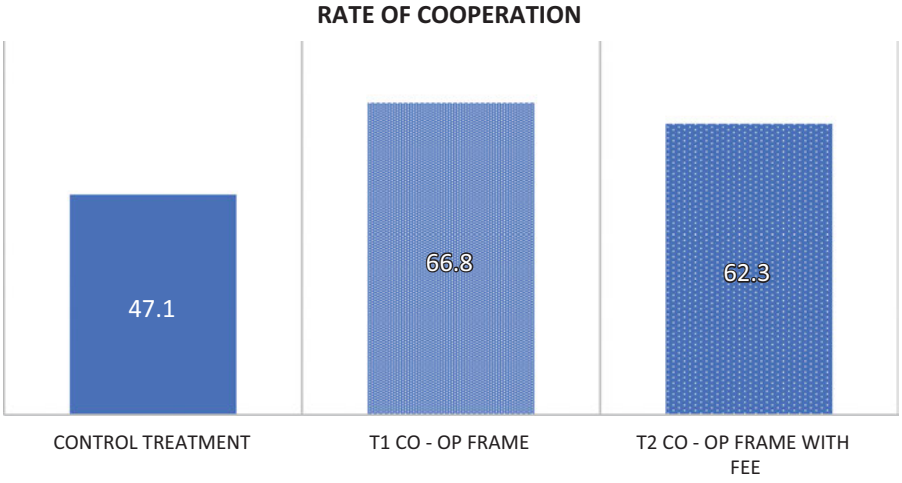


Figure 14.1: Cooperation rates in the framing experiment.

Comparing the absolute amount of group contributions, we see no significant difference between the two treatment groups. The absolute voluntary contribution in T2 is lower, but this is only the case because participants in T2 have a lower endowment (18 ECU instead of 20 ECU). If we add the mandatory fee of 2 ECU from each cooperative member, the absolute average group contribution is almost the same. This means that cooperatives seem to be able to collect the necessary resources

even without mandatory contributions. However, even if mandatory contributions exist, there is no support for the crowding out effects of voluntary cooperation.

Table 14.1 shows that the results are stable when we control for individual characteristics. Only the effect of T2 disappears when we include participants' attitudes towards cooperatives. However, the difference does not disappear. Moreover, we find a significant negative effect of the belief that cooperatives will not work in principle. Interestingly, the effect of T2 ceases to be significant when controlling this belief, whereas the effect of T1 remains stable. Although this could be caused by the small sample size, another possible interpretation is that the mandatory contribution leads to a higher belief that cooperatives do not work (otherwise, a mandatory commitment would not necessary) and that this effect in sum cancels out the positive effect of cooperatives. This idea would be in line with the idea that cooperatives are aligned with the belief that people will be more cooperative, and if this belief is weakened, people lose their trust in the institution.

Table 14.1: OLS regression on cooperation in the framing experiment.

DV: cooperation	Model 1 b/se	Model 5 b/se
Treatment: Ref. CT		
Treatment 1	.26* (0.10)	.24* (0.10)
Treatment 2	.20* (0.09)	.13 (0.09)
Social value orientation	.17 (0.10)	.15 (0.09)
No knowledge about coops		−.00 (0.09)
Coop membership unwanted		−.03 (0.10)
Belief: coops do not work		−.25** (0.09)
Participant has honorary post		−.11 (0.09)
Constant	.06 (0.15)	.27 (0.16)
N = 84		

OLS estimates with robust standard errors; * p < .05; ** p < .01; *** p < .001.
Additional controls are age, sex, income and political party preference.

Although the result of the first experiment is evidence for a general framing effect and thus a normative belief that cooperatives enhance cooperation, the question arises as to whether the decision to obtain a member and the subsequent decision to cooperate is based on those beliefs. This is the subject of the next experiment.

14.4.2 Joining experiment

The second experiment also examines the cooperation-enhancing effect of cooperatives as well as the effects of an entrance fee, which is a sunk cost. However, in this experiment, we additionally observe that the decision to join a cooperative is dependent

on its effectiveness. This allows us to analyze whether people (a) tend to join an institution called a cooperative even if it is not beneficial to do so and (b) the relationship between joining and the cooperation rate.

The experiments took place on study information days in schools and a public event at the university in 2003. As the participants are mainly pupils from the German Gymnasium (high school) and the interested public, we expect no significant knowledge about game theory and cooperation experiments. Participation in the experiment is voluntary and not financially remunerated. A total of 272 persons participated in the experiment.

Again, we use a framed experimental design, this time based on the Classic Chicken Game (see Table 14.2; for similar experiments, see Prosch and Petermann 2001; Prosch 2000). As in the Prisoner's Dilemma, cooperation in a Chicken Game is a Pareto-optimal result. However, in contrast to the Prisoner's Dilemma, the result in the Chicken Game is not stable (see Rapoport and Chammah 2016 for a classical analysis). If one actor in the game assumes that the other actor cooperates, he or she has an incentive not to cooperate. If both actors follow this logic, the result is the worst for both and thus collectively inefficient. The game has two Nash equilibria in pure strategies (C/D and D/C) and one in mixed strategies (for our payoffs, the probability for each strategy is p=0.5)

Table 14.2: Chicken game.

		Player 2	
Player 1		**Cooperation**	**Defection**
	Cooperation	3 / 3	2 / 4
	Defection	4 / 2	1 / 1

Each participant was seated alone at a table in a room and received instructions and the decision form. The participants were informed about the fact that they would play several stages. At the beginning of each stage, they were informed about the rules and the possible payoffs. Participants were told that they would play against a computer that simulates a purely rational player. Each participant played three treatments, starting with the control treatment, followed by treatments A and B. The participants were told that they would play multiple stages, but they did not obtain information about the rules until the beginning of each stage. The resulting points of each round were summed up and copied in a certificate, which was handed to participants.

In the first stage of the experiment, the control treatment (CT), the basic Chicken Game without any modification, played a single shot against another anonymous player. This shows us the baseline cooperation rate, which is used as a reference for the following treatment groups.

In the second stage of the experiment, the first treatment (TA) is introduced. The participants were offered the opportunity to join a cooperative called "Exclusive". In joining the cooperative, the participant automatically makes a costly investment by paying a fee of two points. Additionally, the new member of the cooperative commits to following the cooperative rules. These rules state that a member has to cooperate if the partner is also a member of the cooperative. If the partner is not a member of the cooperative, one has to defect. Not meeting these rules is sanctioned with a fine of two points. The participants knew whether their partner was a member of the cooperative but had no further knowledge about their partner (stranger matching). Then, the basic game is played once. If a player decides to join, the dominant strategy against other members is cooperation and against a nonmember is defection. If a player does not join and plays with a member, the best answer against the member's dominant strategy D is to cooperate. If both members did not join, they enter the basic Chicken Game, without any dominant strategies. It is important to note that joining the cooperative is not a rational strategy: due to the entrance fee, two members, each playing C, obtain only one point, which is the lowest payoff in equilibria possible. Playing the dominant strategy when paired against a nonmember or the best answer as a nonmember against a member ensures two points, and the mixed strategies even lead to an expected payoff of 2.5 points. If players become a member in this treatment, this can be seen as an indicator for normative beliefs that one should join a cooperative since it ensures a cooperative solution (even if this is not efficient).

In the third stage, participants face a second treatment (TB). The participants are again asked to decide if they want to join a cooperative called "Free". The treatment resembles TA, with the exception that in TB joining, the cooperative is free of charge. The other characteristics of the game (one shot with stranger matching) remain the same, including the commitment to cooperative behavior when meeting another member and noncooperative behavior. This rule is again backed up by a sanction of two points. The structure of dominant strategies is the same as in TB: a rational member plays C against another member and D against a nonmember. A nonmember plays D against a member, and nonmembers play a mixed strategy against each other. However, in this case, joining the cooperative is the best option: it ensures three points against other members or even four points against nonmembers, whereas nonmembers can expect two points against members or 2.5 points against other nonmembers.

This structure of treatments sheds light on the normative basis of cooperatives. First, the comparison between treatments TA and TB shows to what degree the participants choose an inefficient institution, which is framed as a cooperative, compared to a completely efficient institution. An explanation for this behavior can be the normative beliefs regarding cooperatives and their results: they ensure a cooperative result under certain costs. Second, the comparison between TA and TB additionally reveals the share of persons who are opposed to the idea of a cooperative: they do not join the institution even if doing so leads to a lower payoff.

Table 14.3 shows the average cooperation rate for each treatment. In the control treatment, that is, the basic chicken game, 63% of participants cooperate. This result is in line with the typical results in similar studies. Having the choice to join the "Exclusive" cooperative by paying a joining fee of two points, 46% of the participants decide to do so, even if joining is inefficient. The majority, however, do not join the cooperative, avoiding the accompanying costs. This could indicate that within the cooperative, we have a positive selection of cooperative and intrinsically motivated persons, who ascribe more importance to the cooperative values than to financial efficiency. However, if we compare the cooperation rates in the base treatment of the participants who will join the costly cooperative in the next step (61.60%) with those participants who will not join the costly cooperative (63.95%), we do not see any difference. The same applies for the comparison of cooperation rates in the base treatment for participants who join or do not join the free cooperative. This means that the selection of people joining the institution is not driven by a general attitude to behave cooperatively but by a kind of preference for cooperative results, even if those are Pareto-inferior.

Table 14.3: Cooperation rates in the commitment experiment; N = 272.

Base Game (CT): 63%			
"Exclusive" Cooperative (TA)			
Ego becomes member (46% of participants)		**Ego does not become member (54% of participants)**	
Partner is member	**Partner is not member**	**Partner is member**	**Partner is not member**
90%	11%	63%	64%
"Free" Cooperative (TB)			
Ego becomes member (81% of participants)		**Ego does not become member (19% of participants)**	
Partner is member	**Partner is not member**	**Partner is member**	**Partner is not member**
96%	16%	53%	71%

If members of the cooperative are paired with another member, 90% of them show cooperative behavior. In contrast, if members meet a nonmember, cooperation rates decrease to only 11%. These results are not surprising since they mirror the dominant strategies in the respective games. Nevertheless, commitment devices considerably enhance the cooperation rates. Participants who decided not to join the cooperative show similar cooperation rates as those in the control group (approximately 63%). However, an interesting point is that nearly one-half of the participants (46%) joined an institution that is clearly inefficient since both members could do better as nonmembers.

The second treatment shows joining rates to the "Free" cooperative of 81%. This is not surprising considering that the decision is efficient. The cooperation rates increase slightly for cooperative members (from 90 to 96%). In addition, 19% did not join the cooperative even if it was inefficient.

Finally, we see that the rate of cooperative behavior within the population is increasing with each step: in total, 171 participants (63%) choose cooperation in the base game, 205 persons (75%) choose cooperation in TA, and 245 persons (90%) choose cooperation in TB. Hence, each institution increases the rate of cooperation within the group, and the effect of the inefficient cooperative is nearly half as large as that of the efficient one.

14.5 Conclusion and outlook

Cooperatives are an often overlooked but nevertheless prevalent form of economic organization (Altman 2010). However, in theory, cooperatives are especially jeopardized by collective good problems since the pooling of resources enables free-riding among members. Consequently, cooperatives need specific cooperation mechanisms to overcome this collective good problem. We discussed two important mechanisms: first, normative beliefs about appropriate behavior in a cooperative, and second, norms that are triggered by joining a cooperative. Based on two distinct experiments, we are able to show that the framing of a Public Good Game as resource pooling within a cooperative leads to higher cooperation rates. This is in line with the finding that similar cooperative frameworks enhance cooperation in Public Good Games (Dufwenberg, Gächter, and Hennig-Schmidt 2011; Ellingsen et al. 2012). In a second experiment, we employed a Chicken Game to analyze the decision to become a member of a cooperative and the resulting effects. Unsurprisingly, we find, in line with a great deal of literature on cooperative behavior, that selective incentives – here, sanctions for noncooperative members – lead to higher cooperation rates. More interestingly, our results confirm that cooperatives seem to have a value on their own: a substantial number of them – nearly one-half of our respondents – joined an institution framed as a cooperative, even if this yielded lower payoffs than staying a nonmember. Within the cooperative, the sanction mechanism leads to a higher cooperation rate.

To our knowledge, this paper is the first to show these framing effects for cooperatives. The results shed light on the institutions in general, which tend to be more effective if they are based not only on positive and negative incentives but also on a normative basis. It is an interesting question as to why cooperatives are especially successful in producing such a basis, whereas state institutions seem to be in need of behavioral incentives to a much higher extent. We can only speculate here, but there are two main differences between cooperatives and many other institutions. First, membership in cooperatives is always voluntary, which allows for a strong

self-selection of members that share the normative basis. Second, cooperatives are based on a high extent of equality and equity among members. These findings may support the evolution of the norms and beliefs regarding cooperatives over time.

We think that it is noteworthy that the results are stable across two different games (Public Good and Chicken Game) and two sampling procedures (economic lab with remuneration and paper and pencil without financial incentives). This makes us confident that the results are stable. Of course, the experiments have a weak spot in regard to external validity; hence, future research should focus on cooperative behavior in real cooperatives.

Appendix: Information treatment for cooperatives in the framing experiment

General information

- The experiment lasts one round.
- The participants of the experiment will be divided into groups of 4 persons each. Each of these groups forms a cooperative. The other members of your cooperative are anonymous, this is, you will not find out who is a member of your cooperative either during or after the experiment.
- Your payout depends on your own decisions and the decisions of the members in your cooperative.

Information on cooperatives

The basic idea of a cooperative is to solve problems together. In contrast to purely profit-maximizing business enterprises, the aim of cooperatives is to provide maximum support for their members. This means that any surplus earned is distributed to the members.

In addition to the economic situation, the social concerns of the members are particularly important. Cooperative values such as solidarity, self-help and mutual support as well as self-administration and self-responsibility determine the actions of their members.

Within the framework of a voluntary and open membership in the cooperative, the members support each other. With the sense of solidarity and self-help, they strive to together improve their economic and social situations. Each individual member of a cooperative benefits from the collectively provided services and resources.

Cooperatives have established themselves in different economic sectors, for example, in the housing sector. In addition to rental and owner-occupied housing, cooperative housing is a widespread form of housing in Germany. A housing cooperative manages its own real estate, maintains it and takes care of renting it out. The members of a housing cooperative share a common interest in affordable housing in which it is worth living. The cooperative union helps the members of a cooperative improve their living conditions, which would not be realizable as an individual person.

Cooperative membership is tied to a one-off payment upon entry. With the help of contribution payments, the cooperative can provide a safe living space at socially responsible prices. In addition, members are free to make an additional financial contribution to support the cooperative.

Literature

Akerlof, George A. 1970. "The Market for "Lemons": Quality Uncertainty and the Market Mechanism." *The Quarterly Journal of Economics* 84 (3): 488–500. https://doi.org/10.2307/1879431.

Altman, Morris. 2010. "Cooperatives, History and Theories of." In *International Encyclopedia of Civil Society*, edited by Helmut K. Anheier and Stefan Toepler,563–70. New York, NY: Springer.

Bonus, Holger. 1994. *Das Selbstverständnis moderner Genossenschaften: Rückbindung von Kreditgenossenschaften an ihre Mitglieder*. Tübingen: Mohr.

Camerer, Colin F. 2003. *Behavioral Game Theory: Experiments in Strategic Interaction*. The Roundtable Series in Behavioral Economics. New Jersey: Princeton University Press. https://ebookcentral.proquest.com/lib/gbv/detail.action?docID=765287.

Cocolina, Carmen Quintana. 2016. "*Cooperatives Europe's report Cooperative Europe key figures 2015*." Accessed August 28, 2019. https://coopseurope.coop/sites/default/files/The%20power%20of%20Cooperation%20-%20Cooperatives%20Europe%20key%20statistics%202015.pdf.

Coleman, James Samuel. 1990. *Foundations of social theory*. Cambridge, Mass. Belknap Press of Harvard Univ. Press.

De Moor, Tine. 2008. "The Silent Revolution: A New Perspective on the Emergence of Commons, Guilds, and Other Forms of Corporate Collective Action in Western Europe". The International Review of Social History 53 (16): 179–212. doi:10.1017/S0020859008003660

Draheim, Georg. 1955. *Die Genossenschaft als Unternehmungstyp*. 2. durchges. Aufl. Göttingen: Vandenhoeck & Rupprecht.

Dufwenberg, Martin, Simon Gächter, and Heike Hennig-Schmidt. 2011. "The framing of games and the psychology of play." *Games and Economic Behavior* 73 (2): 459–78. https://doi.org/10.1016/j.geb.2011.02.003.

Ellingsen, Tore, Magnus Johannesson, Johanna Mollerstrom, and Sara Munkhammar. 2012. "Social framing effects: Preferences or beliefs?" *Games and Economic Behavior* 76 (1): 117–30. https://doi.org/10.1016/j.geb.2012.05.007.

Fehl, Katrin, Daniel J. van der Post, and Dirk Semmann. 2011. "Co-evolution of behaviour and social network structure promotes human cooperation." *Ecology Letters* 14 (6): 546–51. https://doi.org/10.1111/j.1461-0248.2011.01615.x.

Fehr, Ernst, and Urs Fischbacher. 2004. "Social Norms and Human Cooperation." *Trends in cognitive sciences* 8 (4): 185–90. https://doi.org/10.1016/j.tics.2004.02.007.

Fehr, Ernst, and Simon Gächter. 2000. "Cooperation and Punishment in Public Goods Experiments." *American Economic Review* 90 (4): 980–94. https://doi.org/10.1257/aer.90.4.980.

Fischbacher, Urs. 2007. "Z-Tree: Zurich Toolbox for Ready-Made Economic Experiments." *Experimental economics: a journal of the Economic Science Association* 10 (2): 171–78.

Greiner, Ben. 2004. "An Online Recruitment System for Economic Experiments." In *Forschung und wissenschaftliches Rechnen: Beiträge zum Heinz-Billing-Preis 2003*, edited by Kurt Kremer and Volker Macho, 79–93. Göttingen: Gesellschaft für wissenschaftliche Datenverarbeitung.

Harbrecht, Wolfgang, ed. 2000. *50 Jahre Forschungsinstitut für Genossenschaftswesen an der Universität Erlangen-Nürnberg: Beiträge zu den Festveranstaltungen am 21. und 22. Oktober 1999 in Nürnberg*. Veranstaltungen / Forschungsinstitut für Genossenschaftswesen an der Universität Erlangen-Nürnberg Bd. 18. Nürnberg: Forschungsinst. für Genossenschaftswesen.

Hardin, Garrett. 1968. *The tragedy of the commons*. New York: Macmillan.

Hechter, Michael, Karl-Dieter Opp, and Reinhard Wippler, eds. 1990. *Social Institutions: Their Emergence, Maintenance and Effects*. Sociology and economics. Berlin: de Gruyter.

Higl, Michael. 2008. *Theorie der Genossenschaft: Eine industrieökonomische Analyse*. Europäische Hochschulschriften. Reihe 5, Volks- und Betriebswirtschaft 3290. Frankfurt a.M., Bern: P. Lang.

ICA. 2018. "*International co-operative alliance: Co-operative identity, values & principles*." Accessed June 06, 2018. https://www.ica.coop/en/cooperatives/cooperative-identity.

Liberman, Varda, Steven M. Samuels, and Lee Ross. 2004. "The Name of the Game: Predictive Power of Reputations Versus Situational Labels in Determining Prisoner's Dilemma Game Moves." *Personality & social psychology bulletin* 30 (9): 1175–85. https://doi.org/10.1177/0146167204264004.

Marwell, Gerald, and Ruth E. Ames. 1979. "Experiments on the Provision of Public Goods. I. Resources, Interest, Group Size, and the Free-Rider Problem." *American Journal of Sociology* 84 (6): 1335–60. https://doi.org/10.1086/226937.

North, Douglass Cecil. 1990. *Institutions, Institutional Change, and Economic Performance*. Political economy of institutions and decisions. Cambridge: Cambridge University Press.

Ole Borgen, Svein. 2001. "Identification as a Trust-Generating Mechanism in Cooperatives." *Ann Public & Coop Econ* 72 (2): 209–28. https://doi.org/10.1111/1467-8292.00165.

Olson, Mancur. 1965. *The Logic of Collective Action: Public Goods and the Theory of Groups*. Harvard economic studies 124. Cambridge, Mass. Harvard Univ. Press.

Prosch, Bernhard. 2000. "Kooperation durch die Schaffung von Institutionen." In *Normen und Institutionen: Entstehung und Wirkungen*, edited by Regina Metze, Kurt Mühler, and Karl-Dieter Opp, 93–114. Leipziger soziologische Studien 2. Leipzig: Leipziger Uni-Vlg.

Prosch, Bernhard, and Sören Petermann. 2001. "Clubmitglieder, Zuckerbrot und Peitsche." In *Jahrbuch für Handlungs- und Entscheidungstheorie: Folge 1/2001*, edited by Ulrich Druwe, Volker Kunz, and Thomas Plümper, 107–128. Wiesbaden: VS Verlag für Sozialwissenschaften.

Rapoport, Anatol, and Albert M. Chammah. 2016. "The Game of Chicken." *American Behavioral Scientist* 10 (3): 10–28. https://doi.org/10.1177/000276426601000303.

Raub, Werner. 2004. "Hostage Posting as a Mechanism of Trust." *Rationality and Society* 16 (3): 319–65. https://doi.org/10.1177/1043463104044682.

Raub, Werner, Vincent Buskens, and Rense Corten. 2015. "Social Dilemmas and Cooperation." In *Handbuch Modellbildung Und Simulation in Den Sozialwissenschaften*, edited by Norman Braun and Nicole J. Saam, 579–94. Wiesbaden: Springer VS.

Raub, Werner, Vincent Buskens, and Vincenz Frey. 2013. "The rationality of social structure: Cooperation in social dilemmas through investments in and returns on social capital." *Social Networks* 35 (4): 720–32. https://doi.org/10.1016/j.socnet.2013.05.006.

Raub, Werner, and Thomas Voss. 1986. "Die Sozialstruktur der Kooperation rationaler Egoisten: Zur "utilitarist." Erklärung sozialer Ordnung." *Zeitschrift für Soziologie: ZfS* 15 (5): 309–23.

Raub, Werner, and Jeroen Weesie. 1990. "Reputation and Efficiency in Social Interactions: An Example of Network Effects." *American Journal of Sociology* 96 (3): 626–54. https://doi.org/10.1086/229574.

Spear, Roger. 2000. "The Co-operative Advantage." *Ann Public & Coop Econ* 71 (4): 507–23. https://doi.org/10.1111/1467-8292.00151.

Williamson, Oliver E. 1985. *The Economic Institutions of Capitalism: Firms, Markets, Relational Contracting*. New York, NY: The Free Press.

Zerche, Jürgen, Ingrid Schmale, and Johannes Blome-Drees. 1998. Einführung in die Genossenschaftslehre: Genossenschaftstheorie und Genossenschaftsmanagement. München, Wien: Oldenbourg.

Hartmut Esser

15 Rational Choice or Framing? Two Approaches to Explain the Patterns in the Fehr-Gächter-Experiments on Cooperation and Punishment in the Contribution to Public Goods

Abstract: The paper "Cooperation and Punishment in Public Goods Experiments" by Fehr and Gächter from 1999 was a milestone for the change of RCT from its standard versions to the adoption of elements from non-economic fields. The contribution investigates the scope of possible explanations for the observed patterns in the F&G-experiments by Rational Choice Theory (RCT) extended by motives of reciprocity, with the model of frame selection (MFS). Main result is that most findings can be reconstructed rather easily by both approaches – with one exception: After starting with punishment and after withdrawal of this option after 10 rounds subjects should following RCT react immediately with at least some defection, following MFS, however, with keeping a high level of cooperation, independently of motives of subjects. An independent empirical test with data also from other experiments (Hermann et al. 2008) confirmed this hypothesis: no change in cooperation, not even by egoists. Alternative RCT-explanations aiming to find cooperative equilibria for keeping cooperation unchanged by egoists could be the assumption of reputation-effects in finite iterated games. This interpretation, however, seems to be not plausible: Fehr and Gächter tried to control explicitly for reputation effects for all versions, and at least for the stranger-version no reputation effects are expected by RCT. The effect, however, appeared in both versions, and for the stranger-version of the data set by Hermann et al. even stronger than in the original experiment.

15.1 Prologue

There is a long history of consent with regard to basic assumptions on sociological theory and empirical research. It all started when Raub and Voss visited, apparently intentionally, various lectures and seminars across the newly founded universities in the Ruhr area. It was an extremely exciting time, in which the concept now known as 'model of sociological explanation' was developed – around the mid-seventies and thus long before the well-known Coleman's 'bathtub' became popular (see Coleman 1986;

Hartmut Esser, University of Mannheim

https://doi.org/10.1515/9783110647495-015

Raub and Voss 1981, 2017). From the beginning for many researchers, but certainly not for all, the Rational Choice Theory (RCT), which was either unknown or frowned upon within the field of sociology, became the center of the necessary microfoundation of sociological explanation. This included its particularly promising applications to social situations within game theory. Despite all advantages as compared to all other approaches, doubts occurred that it possibly doesn't capture all important aspects of (social) action, for example emotions, internalized norms, habits and – last not least – the 'definition of the situation' by 'significant' symbols, which then defines what has to be considered as 'rational'. Some of those who played a direct and leading role in these developments had, therefore, looked for ways to meet this incompleteness. Examples are Viktor Vanberg who considered rational choice as a respectively activated 'program' within situations (Vanberg 2002), or Siegwart Lindenberg and his to date further and further extended concept of goal framing (Lindenberg 2015). Such relapses had to appear strange to those who were happy to have finally discovered RCT and game theory as instruments, which seemed to make a serious, explaining sociology possible.

The author of this contribution had felt the same at first and in fact for a long time: not again all this drivel about culture! And then the attempts to integrate it into the framework of RCT! Was there any alternative? But there was more and more evidence indicating that RCT wouldn't possibly be the last word and that one wasn't able to cope with various anomalies by – more or less skillful – extensions of the theory. My personal re-framing started with different findings in the course of a long-term project on divorces. Here, *empirical* evidence had shown that – inter alia – the fact of the *ritual* of church wedding *alone* reduces dramatically the probability of a later divorce. This applied even after controlling for really all conceivable RCT hypotheses on this topic, like public commitment, religiosity, or normative climate within peer groups. Moreover, later marital crises are clearly reduced by church wedding, A parallel running qualitative panel study with detailed, intensive interviews revealed that in approximately two thirds of the cases there were virtually no variations in the 'framing' of the relationships, even if extreme stress occurred. In short: If the 'frame' of the relationships was 'matching' from the very beginning and if it was symbolically reinforced, the marriage would be considered as naturally intact, despite all occurring highs and lows, and a separation would be completely out of question.

On the basis of earlier considerations on the integration of approaches of symbolic interactionism for the 'definition of the situation', the theory of everyday behavior by Alfred Schütz, (1971) above all Herbert Simon's concept of 'bounded rationality' (1983) and Thomas S. Schelling's game theoretical contributions on focal points (1960), and the newly emerging script and schema theories as well as dual process models of cognitive social psychology, gave rise to the development of the model of frame selection (MFS). It allowed for – in the first instance theoretically – identifying causally the 'definition of the situation', which is controlled by symbols, and for specifying conditions for a rational choice to occur in a very simple way. And this all happens within the framework of an also formally specified theory and causal relations.

Nearly nobody liked this model (at first). This is true for all kinds of cultural sociologists, who would lose their methodological dualistic autonomy, but also for adherents of RCT – including Raub and Voss. This becomes unequivocally obvious in several passages in a contribution to a 'polemic', which was published for a similar occasion like this one. One example: "First, it is not necessary to introduce so to say different 'forms of rationality' in order to explain normative action. Second, it is decisive to overcome the restriction to a parametrical concept of action and to take into account explicitly social interactions in processes of the definition of the situation." And subsequently: "The model of frame selection is not suitable for this." (Raub and Voss 2009: 188, translation H.E.). Here is another late reaction to these objections.

15.2 Background

The paper "Cooperation and Punishment in Public Goods Experiments" (Fehr and Gächter 1999; also cited as F&G in the following) is not only a typical example of game-theoretical modeling and empirical testing of theoretically derived hypotheses, but it is also a milestone for the change of RCT from its orthodox and strict economic versions to the adoption of ever more elements from non-economic fields like cognitive psychology, sociology and anthropology. A 'general' theory of action for all social science sub-disciplines seems to be conceivable (cf. Fehr and Gintis 2007: 60f.; Gintis 2007: 15f.) and the process of convergence is actually in full speed (Tutić 2015: 84ff.). Main result of the experiment was that in a public good situation, where traditional RCT expects no contribution at all, subjects not only started with a contribution rate of about 50%, but that with merely announcing the opportunity to punish free riders, contributions made a strong jump upwards and converged to nearly 100% in the following rounds – although punishment was expensive for punishers and should not occur for selfish rational actors.

Rational Choice Theorists and Behavioral Economists usually are convinced that an explanation of these and other deviations from traditional RCT is possible with rather minor changes in the two core elements, which drive the rational selection of actions: preferences and beliefs, and that it is *not* necessary (or even a bad mistake) to change the *general* micro-foundation of the selection of action as future oriented, strategic 'choice' itself. That means that other processes to explain action and behavior like pattern recognition, categorization and the triggering of reaction-programs by (subtle) symbolic cues in a situation as they were found in the seminal experiments by Tversky and Kahneman (1981) on the framing of decisions by different wordings for same incentives or by Liberman et al. (2004) on contributions to collective goods by labeling a situation with same incentives as 'Community' or as 'Wallstreet' Game were unnecessary. In both cases it became evident, that peculiar preferences to cooperate or to defect seemingly play no role. Irrelevance of other

aspects than preferences and stability of preferences, however, are two of the key elements of any version of RCT. The findings give thus good reason to take up this point: Is it possible to explain the patterns in the F&G-Experiment by means of RCT keeping the principles of utility maximization untouched? Or is it necessary to change not only assumptions about preferences, beliefs and utility functions but also the mechanism for the micro-foundation in face of such results.

The aim here is to compare the story RCT has to tell to explain the results of the F&G-Experiments with the story of the MFS, which has been considered by many researchers as being not suitable at all to deal with social situations and the dynamics of interactions. The following describes first the structure and core results of the F&G-experiment in its original version from 1999, gives a short overview of the core elements of RCT and MFS in comparison, tells two (short) stories to explain the crucial patterns of the findings with a summary for the fit of theoretical predictions with the findings, and presents an additional empirical test of both approaches.

15.3 The experiment

The peculiar aim of the F&G-experiments was a systematic test for effects of opportunities to punish free riders in public good games. The prediction following strict orthodox RCT was that nobody would contribute and nobody would punish if punishment is costly. With a tiny bit of incomplete information cooperative equilibria are possible, but the specific patterns of the experiments hardly can be explained with egoistic motives alone. Fehr and Gächter therefore assumed "reciprocity" as an additional motive – over and above of altruism, which could explain cooperation, but not (costly) punishment, because punishment reduces the gains of altruism (see also section 6). The experiment lasted 20 rounds. It was split up after 10 rounds with one of the two opportunities to change to the other version for another 10 rounds. Subjects were informed about the contributions of the others after every round and knew that the experiment would be terminated after round 20. Experiments were varied in two different ways: Without and with punishment opportunity first and with randomly varying subjects in the single rounds (stranger condition), or with the same subjects over all 20 rounds (partner condition). The aim was to control for reputation effects and effects of the experience of a common history.

15.4 Main findings

Figure 15.1 shows the pattern of the two variants of the experiments (starting without punishment NoP1 and with punishment WiP1 in the first 10 rounds followed by a reversal in the following 10 rounds WiP2 and NoP2, respectively, each in the partner version).

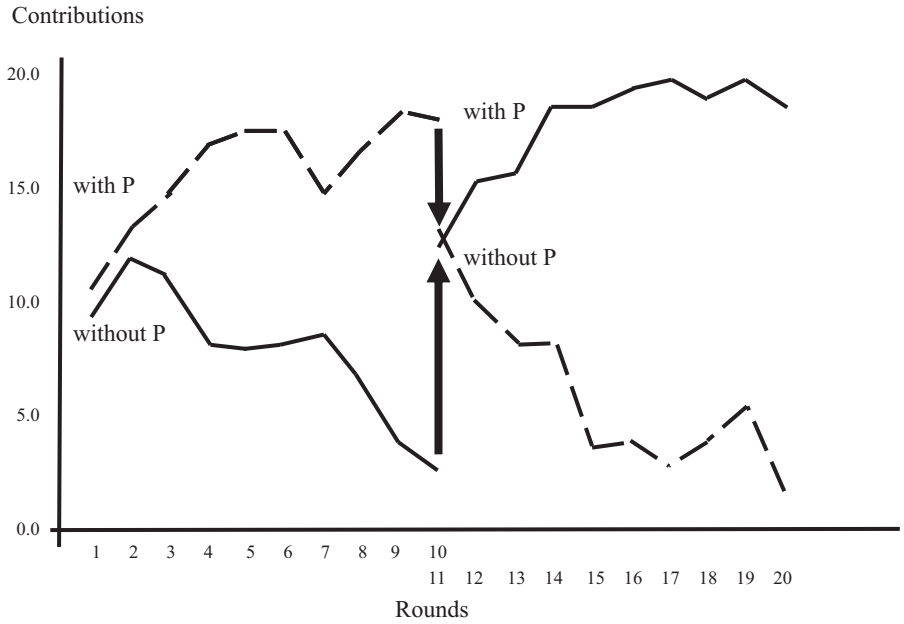


Figure 15.1: Patterns of contributions for two variants in the F&G-experiment (partner version).

The findings can be summarized in 6 observations.[1]

1. With no punishment option in round 1 (NoP1) an average of about 10 units (or 50% of the maximum of 20) of contribution to the public good is observed.
2. Contributions deteriorate in the following rounds more or less steadily and approach the level of nearly overall defection in round 10.
3. With the announcement of a punishment option after round 10 (WiP2) contributions rise in round 11 immediately in a strong jump of about 10 units.
4. After round 11 with the punishment option a steady increase of contributions takes place up to about 19 of possible 20 units.
5. The general shape of the patterns is similar for both variants of order: convergence of cooperation up to nearly 100% and decay with WiP down to nearly full defection with NoP.
6. If punishment comes first (WiP1), a rather small decrease of about 4 units (from about 18 down to about 14) takes place from round 11 with no punishment (NoP2) in contrast to the jump up of 10 units up if punishment comes after no punishment (WiP2) and the decay in cooperation (from ca. 3 up to ca. 13).

1 For an extended and more detailed comparison on 14 findings see: Esser 2018: 19f.

These six findings make up the core of the following comparison of RCT and MFS in being able to explain the patterns. The comparison refers to differences between RCT and MFS in their theoretical arguments and their ability to offer explanations without changing basic assumptions.

15.5 RCT and MFS

Before we start with the comparison we summarize shortly the peculiarities of RCT and MFS as background for their theoretical predictions for the outcomes in the F&G-experiment (cf. for a summary of central RCT-assumptions: Gintis 2007, part 4; of main elements of MFS: Esser and Kroneberg 2015; and for a comparison of basic tenets of both approaches Tutić 2015, parts 2 and 3). One reason is that at least the peculiarities of MFS cannot be presupposed for most readers, in- and outside the RCT-camp, but also to demonstrate, where the substantial differences become obvious. Aspects are the basic principles to explain (re-)actions, the modeling of social situations, the explanation of processes and changes in view of each approach. Especially it should become clear, that the MFS cannot be interpreted simply as another extension of RCT – as some representatives of (orthodox and wide) RCT as well as most critics from Interpretative Paradigm or cultural sociology are inclined to do.

Two elements can be assigned as *basic principle* of RCT. First, action is conceived as 'choice' between alternatives by maximizing the product of preferences and beliefs for consequences in the future. Second, choices have to fulfil certain axiomatic conditions like transitivity and independence of variations in their verbal presentation. Situational cues should, therefore, have no systematic effect on preferences, but possibly on beliefs. RCT varies in some variants between fully informed homo oeconomicus with transitive preferences and differing variants of 'bounded' rationality. The respective type of rationality, however, is assumed as fixed and not subject to situational change. Emotions are not part of common versions of RCT, but are sometimes conceived of as special kind of preferences. All effects of incentives, costs or risks are conditional: They add to expected gains or losses, but without changing the respective utility functions. *Social situations* are characterized in RCT by common knowledge about opportunities, beliefs, and preferences of others in a situation and by anticipation of their possible reactions. The main aim of RCT is the derivation of typical equilibria for collective outcomes, in particular by instruments of game theory. *Processes* of strategic interaction are conceived as iterations and causally connected chains of single acts. *Changes* in behavior and collective outcomes are possible by two mechanisms: first, anticipation of reactions and gains/losses for a complete sequence and, second, adaptation of beliefs from the observable behavior of others by (Bayesian) learning, for instance as "conditional"

cooperation or defection. Preferences are assumed to be stable over all situations, but can change in the long run by adaptive reinforcement-learning and evolution.

Basic principle of MFS is a process of decoding 'significant' symbols ('cues') in a situation, which trigger the activation of a specific mental model for the "definition" of the situation, the 'frame', connected programs for reacting, more or less habitualized 'scripts' for sequences of behavior and single acts. Mental models and reaction programs were encoded in the past mainly by socialization, but in part also by biological evolution. They contain typical patterns of main goals, preferences, beliefs and emotions for typical situations. The set of mental models for different situations makes up the (multiple) identity of actors. The strength of the activation for a specific frame depends on the mental model's internal accessibility, the external presence of a particular physical object, the cognitive link between mental model and physical object, and the degree of occasional noise for this link. If the match between object, cue and a (strongly) accessible mental model is perfect, the activation of a frame and the execution of the connected behavior will follow immediately and without any 'choice' and deliberation of consequences in the future. The MFS denotes it as as-mode of selection. Main implication of a strong framing is the suppression of the effects of other incentives, costs, and risks with corresponding changes of the respective utility function up to the complete unconditionality of a frame. In case of a certain mis-match *and* strong motives and available opportunities to deliberate consequences, a 'rational' choice can take place – up to a level even orthodox RCT and *homo oeconomicus* presumes. MFS denotes it as rc-mode of selection. Insofar RCT could be regarded as a special case of the MFS, but with an explicit consideration of a peculiar 'definition' of the situation preceding any 'choice' – what is missing in all versions of common RCT. *Social situations* can be conceived in two ways in the MFS. First, as framing of the situation by mental models for types of common actions and collective outcomes. If framing is strong, a kind of mechanical coordination by the automatic activation of the same frame of social action for all actors will simultaneously take place. Second, if framing is weak and motives and opportunities for deliberating consequences are given, strategic social action will become possible – just in the same way as RCT and game theory conceive social situations – but again with a preceding definition of the type of social situation. *Processes* can be modelled as sequences of single acts of framing and subsequent behavior with effects on the next steps of framing and (re-) action, possibly also with switches between automatic and strategic (re-) actions in an as- or a rc-mode. *Changes* in framing situations (and in the modes of selection) can occur by two mechanisms. First, alterations in decoding and categorization, e.g. by sudden mis-matches and/or the appearance of cues indicating another type of situation. Second, changes in encoding mental models by adaptive learning from deviations with regard to content, accessibility and the link, which transforms objects in 'significant' symbols. For both changes actions of others can serve as cues for the framing of the situation in the next rounds.

Table 15.1 summarizes the core constructs, proposed mechanisms and additional assumptions for both RCT and MFS (as-mode only).

Table 15.1: Basic constructs and assumptions of RCT and MFS.

Aspects	RCT	MFS
Mechanism	Choice max. expected utility shadow of the future	Categorization (Mis-)Match shadow of the past
Opportunities	yes	yes
Beliefs	yes	yes
Preferences	yes	yes
Emotions	(no)	yes
Cues/Symbols	'cheap talk' relevant for beliefs only	'definition of the situation' relevant for beliefs and preferences
Types of rationality	fixed	variable
Unconditionality	no	yes
Social situations	strategic interaction	mechanical coordination
Processes	Sequences of situations and actions	Sequences of situations and actions
Changes: short-term	Bayesian learning Beliefs only	(Mis-)Match/Re-Framing Beliefs and preferences
long-term	adaptive reinforcement evolution	adaptive reinforcement evolution

15.6 Two (Short) stories

RCT-explanations of the empirical patterns of the findings in the F&G-Experiment (Figure 15.1) start with the assumption of social motives as preferences and two types of actors: egoists (EP) and altruists (AP). In WiP1 both types start with their preferences: EP-types defect, AP-types cooperate. The sharp decay of cooperation after a certain period of stability can be explained by 'conditional cooperation': Altruists learn, that they are exploited and that it is better to defect, simply as defense. With the announcement of the punishment opportunity Altruists have an offensive option: to punish defectors. Cooperators with only altruistic motives, however, would not choose this option, because any punishment is not only costly, but decreases their own benefits, like the feeling of a warm glow by cooperation itself and for the well-being of others. In order to explain the rise of contributions after the decay in round 11,

additional motives have to be assumed: Cooperators are not only altruistic, but 'reciprocals' (RP) with strong motives to retaliate against defections. EP-types know this, anticipate severe punishment and cooperate, because this pays now for them. RP-types anticipate this and start for WiP2 with the punishment option in round 11 with cooperation – on a higher level not only after the decay before, but also in comparison to the beginning in NoP1. The steady rise of contributions after round 11 up to nearly 100% can thus be explained easily: *Both* EP- and RP-types have good reasons to continue with cooperation – even if not all participate immediately, but quickly learn that cooperation is better for them, too.

The MFS-explanation of the pattern starts also with a distinction of motives, but *not* as stable characteristics of *actors* over different situations, but as varying imperatives in *situations*. The no punishment situation in the F&G-Experiment is objectively one of the type of a prisoner's dilemma, but not framed very clearly. If we assume some kind of default options for not clearly defined situations, Cicourel's 'basic rules of interactions', results for round 1 can be explained by following their mental models of EF or AF: 50% cooperation, 50% defection in each case. If reactions in the first round (and later) serve as cues to define the situation more clearly, the decay in cooperation can be explained as stepwise increase of a framing as Wallstreet Game, where selfishness and competition prevail, until all actors follow this definition of the situation regardless of their social motives. The announcement of punishment after round 10 serves in a similar way as significant cue for another definition of the situation: not as Community Game with rather innocent altruism and friendliness, but as 'Reciprocity' Game, connected with strong inclinations to retaliate against defections immediately and without any consideration of further consequences as the costs of punishment. This not only explains the higher level of contributions in round 11 as compared with the undefined situation in round 1 with its random or default reactions, but also the sudden jump of about 50% difference up from the level after the decay of cooperation before. The steady rise up to nearly full cooperation in the last round would be a consequence of the increasing match for the reciprocation frame – backed by effectively punishing defectors and collecting more and more collective gain – up to the end of the last round.

15.7 Comparison

Table 15.2 provides an overview of the theoretical arguments and assumptions of the two approaches to explain the six findings mentioned above.

For findings 1 to 5 both approaches can rather easily explain the respective patterns (with different theoretical arguments). The base is a peculiarity of reciprocity: Common knowledge of the possibility of self-inflicting retaliations by (strong) emotions is presumed by RCT for the assumption of a subgame perfect equilibrium

Table 15.2: Theoretical arguments and assumptions to explain the F&G-findings.[2]

Findings	RCT	MFS
1	Types of preferences (EP, AP)	Types of frames (EF, AF)
2	conditional cooperation: beliefs C/D	Mis-Match AF
3	common knowledge P with WiP2	activation RF with WiP2
4	common knowledge P with WiP2	increasing match for RF
5	similar conditions in both versions	similar conditions in both versions
6	**no explanation**	Priming RF by C in WiP1

given the punishment option, instant activation of strong affects and inclinations to retaliate against violations of cooperation- and fairness-norms as part of the reciprocity frame activated by the announcement of the punishment option is presumed by the MFS. Finding 6 indicates, however, a difference. RCT predicts a clear decrease in contributions after removing the punishment option as first sequence: Selfish defectors now have not longer to fear retaliations, and altruistic cooperators are facing the risk to be exploited. MFS in contrast assumes a process of strengthening accessibility and match of the reciprocation frame first with a following history of high, stable and unbroken cooperation: An increased 'priming' of AF by repeated cooperation of nearly all. Defection after removing the punishment options could, therefore, be perceived as occasional mis-match of a firmly established frame in which cooperation prevails. This explains the rather small decay of contributions with the change from punishment to no punishment – in contrast to the reversed situation of the decay of cooperation first, increasing anger of exploited cooperators, who will change immediately to reciprocation if punishment becomes possible. Finding 6, however, confirms MFS, but RCT has *no* explanation for the asymmetry in changes of contributions between round 10 and 11 for the two variants of NoP1 and WiP1 (highlighted in Table 15.3).

The summary in Table 15.3 refers to the most extended variant of wide RCT so far: from the orthodox version with the assumption of selfish motives only (RCT 1.0) over the extension by motives of conditional cooperation, warm glow, altruism, fairness or inequity aversion (RCT 2.0) to the assumption of another type: strong reciprocation with a nearly unconditional willingness to punish violations of group standards (RCT 3.0). Table 15.3 summarizes how the different versions of RCT fit with the 6 findings (numbers indicate confirmation of the resp. finding).

2 Preferences: Egoism (EP), Altruism (AP), Reciprocity (RP); Frames: Egoism (EF), Altruism (RP), Reciprocity (RP); (Re-)Actions: Cooperation (C), Defection (D), Punishment (P).

Table 15.3: Theoretical predictions and empirical fit between variants of RCT and MFS.

		theoretical predictions and empirical fit				
		RCT			MFS	
F.	Content	RCT 1.0	RCT 2.0	RCT 3.0	PBB	RCT*
1	50% C rd. 1 for NoP1	refuted	1	1	1	(1)
2	15% C rd. 10 for NoP1	2	2	2	2	(2)
3	65% C rd. 11 for WiP2	refuted	refuted	3	3	(3)
4	95% C rd. 20 for WiP2	refuted	refuted	4	4	(4)
5	same both versions	5	5	5	5	(5)
6	asymmetric change rd. 11	not expl.	not expl.	not expl.	6	(not expl.)

RCT 1.0 is clearly refuted by most of the findings. Only two of the six findings are in accordance with selfish utility maximization in public goods games: decay of cooperation as consequence of attempts to defend oneself against further exploitation (finding 2), independently of the order of both sequences (finding 5). The assumption of altruism as social motive in RCT 2.0 allows an explanation of even high rates of cooperation in the first rounds of the sequences (finding 1) and for the larger effects for partners (finding 9): Motives for cooperation compensate probably the selfish ones, and for cooperation with partners higher amounts of warm glue and stronger beliefs in the reliability of their type may arise. To explain the effects of the punishment option, however, more than social motives of warm glow or altruism has to be assumed: credible willingness to strongly and unconditionally retaliate against defective deviations from a group standard (RCT 3.0). With one exception, all findings can be explained with these assumptions. The exception concerns the rather small decrease of cooperation after punishment: Defectors don't change their identity even by long sequences of successful cooperation before in round 11 for WiP1 to NoP2, and they should reduce their contributions immediately just by the *same* amount, by which they increased their contributions with announcement of the punishment option in round 11 for NoP1 to WiP2.

The listed arguments of MFS in Table 15.2 refer to the as-mode of an automatic activation of a certain kind of what Victor Vanberg sometimes has called program based behavior (PBB). All 6 findings fit with the MFS-predictions for this case. MFS, however, comprises also 'rational choice' as selection-mechanism – provided that conditions for the rc-mode are met and, thus, in principle for all versions of RCT as special cases (including the applicability of all formalizations and instruments like game theory). But this does not mean a straightforward subsuming of common RCT simply under the rc-mode of the MFS: For the MFS, *all* selections are *preceded* by a

process of *categorization* and a certain 'definition' of a situation, and it depends on the *strength* and *content* of framing, what happens. The strength of framing determines the degree of (variable) rationality and the content how a situation is seen and 'interpreted' by actors. Framing, then, can be a matter of automatic activation as well as of a 'rational choice'. This implies that any 'definition' of a situation will lose its unconditionality, if cues are ambiguous and accessibility for a specific mental model is weak. In this case other motives can interfere, like especially selfishness, which always tends to be pushed to the fore and to override social motives and the impulse for costly emotional reactions (according to Goal Framing Theory by Lindenberg 2015: 47ff.). This specific MFS-version of RCT *with* framing is labeled as RCT* – in remembrance of the concept of 'unit act' proposed by Talcott Parsons already in 1937, who postulated that *any* act is preceded by a 'normative orientation', which *defines* the situation with respect to preferences, beliefs and perceived opportunities. The MFS is the formalized version of this concept, enriched and completed by well-established insights and findings from cognitive psychology, neuroscience, anthropology and evolutionary biology since this contribution of Talcott Parsons to a 'General Theory of Action'.

For the F&G-Experiment RCT* would make in principle the same predictions as version RCT 3.0 (with the same arguments, assumptions, confirmations, refutations, and blanks in explanations). But RCT 3.0 assumes a preceding step of 'definition' of the situation with effects like those in the experiments by Tversky and Kahneman or Liberman et al. That leads to another MFS-story – now for framing the situation in the *rc*-mode. Beginning again with no punishment first and assuming a rather open, undefined situation (as in the F&G-experiment) selfish or social preferences and types of actors determine what happens: More than 50% cooperation from the beginning, conditional cooperation and decay. Assuming a clear cue for cooperation- or reciprocity-frame contributions should change: an even higher cooperation rate and no decay. Wallstreet-cues would have the reverse effect: suppression of cooperation and no change for the following (as in the Liberman-experiment). For all effects and patterns possible framing-effects, however, would be weaker in case of a rc-mode as compared to an as-activation: A certain mismatch is one of the necessary conditions for any (rational) deliberation. And referring once more to Lindenberg's Goal-Framing Theory, selfish motives should become stronger with the degree of (possible) deliberation. Exactly that is observed accidently and mostly more as by-catch in other experiments of behavioral economics (Costa et al. 2014; Duwfenberg et al. 2011; Ellingsen et al. 2012; Engel and Rand 2014; Rand et al. 2012; Rubinstein 2013). In Table 15.3 predictions and fits with the 6 findings are, therefore, presented in brackets for version RCT* – 'Rational Choice' in the rc-mode of the MFS.

15.8 A separate test

The only case where the approaches contradict each other theoretically is finding number 6 with the asymmetry occurring when the punishment option is changed between round 10 and 11. This relates directly to the core of both approaches: According to RCT, *same* reactions should occur both upwards and downwards, regardless of whether the punishment option is introduced or abolished in the first or second sequence. The *only* thing that counts is the common knowledge of the options, and neither introducing or abolishing the punishment option leads to changes in preferences. By contrast, according to the MFS on would expect *different* reactions (in the as-mode): Introducing the punishment option after prior decay of cooperation without punishment option (transition NoP1 to WiP2) will result in a significant jump upwards, too. Abolishing the punishment option after having started with it and after the subsequent sequence with high cooperation rates (transition WiP1 to NoP2), however, will (at least in the first instance) will cause *no* special change. The theoretical argument for this refers to the basic mechanism of frame selection: the increasing match of a cooperation (resp. reciprocity) frame via a process of an ever stronger activation of the framing, being reinforced by cooperative acts themselves, the iterative symbolic 'constitution' of a cooperation 'community', just as it had been observed in Liberman's experiments in the 'community game' or analogous field experiments on priming of altruistic attitudes (cf., *inter alia*, Keizer et al. 2013).

So much on the reconstruction of the findings that can be reconstructed from the publication of the F&G experiments. But would there also be a chance for a test independently of the published results? An obvious way is linked to the core of the respective approaches: actors' preferences (EP and RP) and the different frames (EF and RF). It is possible to derive specific hypotheses from both approaches with regard to changes in the transition from round 10 to round 11 for both versions (NoP1-WiP2 and WiP1-NoP2).

RCT suggests that introducing or abolishing the punishment option *alone* will change the respective expectations: Accordingly, changes would amount to the *same* effect on cooperation for *both* types in *both* directions (see the left-hand side of Figure 15.2: RCT, NoP1-WiP2 and WiP1-NoP2 for EF and RF). The main reason is that following RCT *beliefs* only would change, but preferences should remain constant. If beliefs on other´s cooperation are increased, however, temptations to defect after the withdrawal of the punishment-option would be even *higher* than before and rather strong defection should follow in round 11. In contrast, according to MFS one would expect a differential pattern (see the right-hand side of Figure 15.2: MFS, NoP1-WiP2 and WiP1-NoP2 for EF and RF). This is because changes vary with the content of framing (EF and AF type) and with the accessibility of respective cues, here the announcement that a punishment option is introduced or abolished. Four theoretical expectations result from this: *First*, in version NoP1-WiP2 there is a clear change

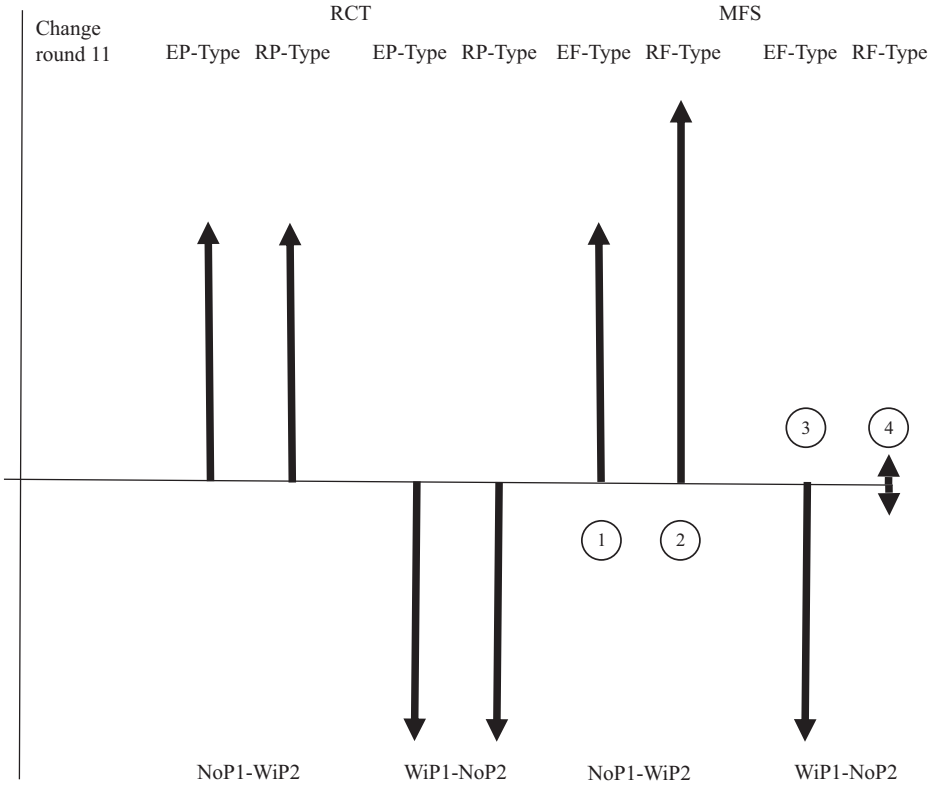


Figure 15.2: Theoretical expectations of RCT and MFS regarding the effects of introducing or abolishing the punishment option between rounds 10 and 11 of both versions NoP1-WiP2 and WiP1-NoP2 according to actors' preferences and frames (EP/EF type and RP/RF type; cf the text above on the different numbers).

upwards for EF types after introducing the punishment option after round 10, quite similar to the magnitude of the change as proposed by RCT (for all types), because their expectations change (Figure 15.2, number 1). *Second*, among RF types cooperation increases clearly, too, in version NoP1-WiP2 in round 11. However, it *exceeds* the changes among EF types (and all of RCT), because the reciprocity frame is (only) available for RF types (Figure 15.2, number 2). *Third*, in version WiP1-NoP2, contributions immediately *decline* among EF types after a long period of high cooperation between rounds 2 and 10: What counts for them (as for RCT in general) are only expectations, and if the risk for punishment no longer exists, they will have *no* reason for cooperation anymore. In addition, there is *no* 'community' priming, because it is not available to them (Figure 15.2, number 3). *Fourth*, according to the MFS, nearly *no* changes will finally occur among RF types in the transition WiP1-NoP2, at least not in round 11 (Figure 15.2, number 4): Among them, the RF frame is ever further activated

through mutual cooperation. Only more obvious and repeated disruptions of the prior 'constitution' of the cooperation 'community' may finally give rise to conditional defection in subsequent rounds.

Three specific hypotheses on the relationships of differences in contributions between rounds 10 and 11 result from this. *First*, the difference in the change of cooperation between rounds 11 to 12 as regards version NoP1-WiP2 is significantly positive for RF types as compared to EF types (compare number 1 with number 2). *Second*, between versions NoP1-WiP2 and WiP1-NoP2 there is no difference with regard to the amount (!) of the change among EF types: Their contributions decrease to the same extent when the punishment option is abolished as it increased before its introduction (compare the amount of number 1 and 3). *Third*, among RF types no change in cooperation occurs in the transition for version WiP1-NoP2 in round 11 as compared to round 10 (number 4).

How could one make an empirical comparison if the types weren't measured separately? One simple consideration provides the solution: The behavior in question when options change, relates to what happens in rounds 10 and 11, but *prior* sequences exist and one can presume at least for round 1 that actors followed (largely) their (private) motives and mental models. These reactions can, therefore, be used as (proxy) measurements of the types, analyses can be conducted separately for both types, and the findings can be compared with the hypotheses stated in Figure 15.2. This is exactly what had been done then.

As the number of cases in the original experiment (Fehr and Gächter 1999 – for Zürich) appeared to be too small for separate analyses, another dataset which is comparable with the approach was used and pooled.[3] The types were dichotomized according to their contributions in the respective first round. All subjects who had spent *at least half* of their available means (10 out of 20 units) were classified as RF types and those who had spent *less* than one half as EF types, respectively. This corresponds to a central theoretical aspect in the explanation of reciprocity effects: Already simple fairness norms include sharing as being part of the basic rules of interaction and only those who stay below can be designated as egoist. This creates a certain asymmetry in the distribution of the types, because there aren't so many

3 The dataset used by Hermann et al. (2008) includes altogether 17 cities from different national cultures. Except the one for St. Gallen, none of the experiments there corresponds to the approach of the original experiment in F&G 1999, because either they considered only one version of introducing the punishment option, or they conducted fewer rounds (for example, six rounds twice instead of 10 rounds twice). Anyway, St. Gallen would have proven more suitable for an analysis, because with it the wider national-cultural context was held constant. Zürich was also included in the complete dataset, but it's structure didn't correspond with the one in the F&G experiment, too. Different from the original experiment, however, only a stranger version of the experiment was conducted, which normally shows the whole pattern less pronounced (cf. the findings 13 and 14 on partner versions in Fehr and Gächter 1999). This makes the test conducted here even more thorough as compared to the hypotheses of the MFS.

egoists then, particularly when the game starts with an available punishment option (WiP1-NoP2). This cannot be avoided, but testing for 'significant' effects is thus conducted even more thoroughly.

The findings of the analyses testing the three hypotheses are summarized in Table 15.4.

Table 15.4: Changes in contributions at the transition from introducing or abolishing the punishment option between rounds 10 and 11 from NoP1 to WiP2 and from WiP1 to NoP2 for egoists and altruists (EF and RF types) using the pooled dataset (Fehr and Gächter 1999 for Zürich; Hermann et al. 2008 for St. Gallen); bold: difference significant with p≤0.001.

	Average Contribution			**Hypotheses MFS (H) and findings F&G (B)**						**n**
	a	**b**	**c**	**1**		**2**		**3**		
Version	**Round 10**	**Round 11**	**Diff. 11–10**	**H**	**B**	**H**	**B**	**H**	**B**	
NoP1-WiP2	3.4	12.3	**8.9**	plus	**4.2**					25
EF Type	2.4	15.5	**13.1**							47
RF Type										
WiP1-NoP2	8.9	8.5	−0.4			zero	**−8.5**	zero	1.4	13
EF Type	15.6	17.0	1.4.							51
RF Type										

The *first* hypothesis is clearly confirmed (column 1): The difference in contributions between rounds 10 and 11 (columns a and b) among RF types exceeds by 4.2 units significantly the one among EF types (column c). In contrast, the *second* hypothesis of the MFS proves to be incorrect, at least at first sight (column 2). It is expected that the amount of change among EF types between rounds 10 and 11 remains the same for both versions, NoP1-WiP2 and WiP1-NoP2, and thus, this difference equals zero (column 2H). Yet, the change during transition to WiP1-NoP2 amounts only to -0.4 (column c WiP1-NoP2, EF type), so that the difference with regard to the change in version NoP1-WiP2 is smaller by 8.5 units that the reference value of the change of 8.9 units for the transition regarding NoP1-WiP2 (column c, NoP1-WiP2, EF type). To put it more simple: Contrary to hypothesis 2 of the MFS, EF types virtually don't respond to the abolishment of the punishment option in round 11 after a long period of prior cooperation. The third hypothesis is then again confirmed clearly (column 3): With regard to version WiP1-NoP2, RF types continue their high levels of cooperation, which indeed even further increase by 1.4 units: from 15.6 units in round 10 to 17.0 units in round 11 (column c, WiP1-NoP2, AF type). The findings are summarized graphically in Figure 15.3.

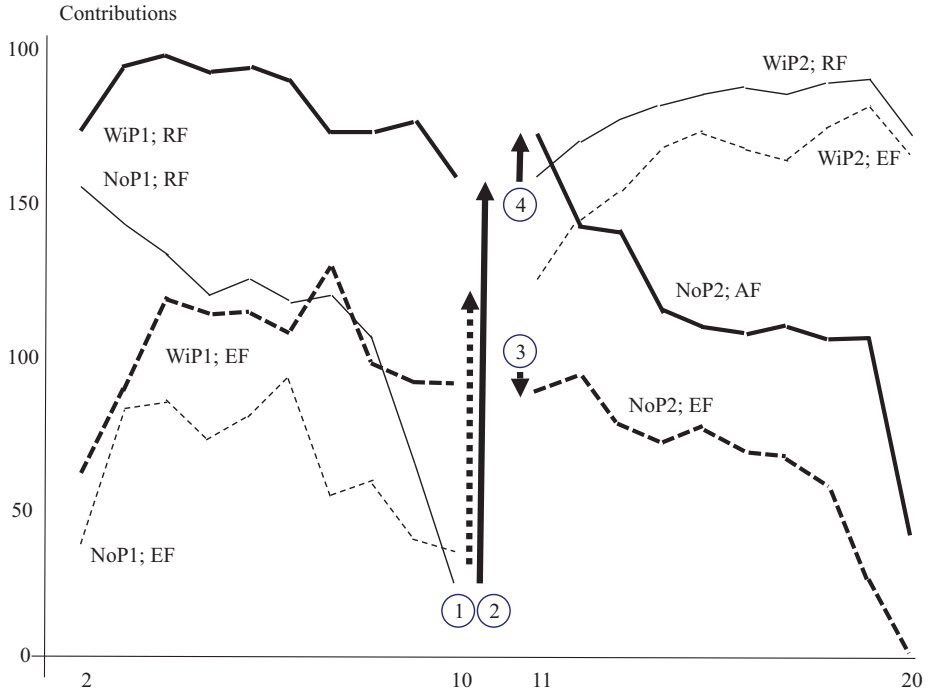


Figure 15.3: Cooperation in the F&G-experiment for the two versions NoP1-WiP2 (thicker lines) and WiP1-NoP2 (thinner lines) in dependence on actors' frames (EF types: dashed lines, RF types: solid lines; cf. the text on arrows and numbers; without round 1, which was used to determine the different types).

The according to hypothesis 2 greater decline in cooperation among EF types (Table 15.4, column 2) looks like a refutation of the MFS. But is this really true? Hypothesis 2 is based on the really extreme assumption that egoists are basically *not* susceptible to a cooperation priming over long sequences of a profitable community. This seems not to be true and there is considerable ethnographic evidence in support of this. Already the various experiments by Liberman had demonstrated that the 'definition' of a game's framing can override *all* preferences. Although differences in the level of cooperation between EF and RF types remain in phase WiP1, the decline in round 11 with NoP2 predicted by RCT (3.0) doesn't occur. Despite all egoism there obviously always exists also a certain susceptibility to processes of symbolic constitution of a cooperation community and a mitigation of reactions to changes in opportunity structures and incentives alone. In any case this finding is a clear rejection of RCT, even in its widest version: namely the assumption that 'framing' and 'definition of the situation' don't exist and that symbols and language are nothing but 'cheap talk'.

15.9 Evaluation

Reconstruction of the F&G-experiments with RCT and MFS shows, that both approaches work rather well. But both have also to make several assumptions without a really sound theoretical and/or empirical foundation until now.

The main problem for RCT is the stepwise extension of types of motives and actors, sometimes appearing simply as post-hoc-'explanation' of clear refutations of the foregoing version of RCT by assuming a new motive for what had just been observed: selfishness, altruism, strong reciprocity, the last including the assumption of the activation of emotions for explaining costly retaliations and thus transcending the boundaries of core elements of common RCT again. It looks sometimes very much like a degenerative problem shift. Two of the six findings could not be explained even by version RCT 3.0 with reciprocity as a stable trait of types of actors, and especially the absence of end-game-effects in the very last round after high cooperation contradicts each of the three versions of RCT.

The MFS is able to explain these two anomalies of RCT 3.0 rather easily if conditions for the as-mode were met: A longer period of cooperation strengthens at least a Cooperation-frame, and single interferences, e.g. by announcement that the punishment option is withdrawn now, are not able to re-frame the situation instantly. The same should apply for end-game-effects if punishment is possible and high cooperation prevails in the second sequence. And: Emotions are not a strange addendum for MFS, but an essential part of any program based behavior (PBB) in an as-mode of framing, activated in an uncontrollable way by certain 'significant' symbols or cues. Indirect evidence for framing-effects, including the activation of emotions and unconditional retaliation of defections, which serve as (very) significant cues triggering aggressive reaction in a Reciprocity-frame, is given by Fehr and Gächter themselves (see above). For RCT* and rc-condition predictions and preceding framing are somewhat different to those of common RCT without any assumption of preceding framing: The 'rational' choices hinge at least partly also on a 'rational' selection of a specific frame, but the effects are weaker in its strength and more selfish in content – the less strong a situation is defined and the more opportunities to deliberate are available. Predictions and fits for RCT* are therefore also marked by brackets in Table 15.3.

The detailed analysis of the F&G-Experiment leaves the question quite open, whether RCT delivers a satisfying explanation within its (more and more opened) boundaries or whether MFS fares better. Both approaches have at least one problem in common: Any additional assumption decreases the logical content of a theory, and that applies to RCT as well as to MFS. RCT has to extend types of motives and actor – up to the assumption of unconditional emotions, and MFS has to distinguish and verify types of frames and certain conditions for the mode-selection. Because both approaches differ strongly in core constructs and assumed mechanisms they can hardly be compared in their logical content. But such an incommensurability is

inevitable in most cases of 'correcting' explanations and theory development. It is the price for achieving a 'comprehensive', general theory, which explains the anomalies of different approaches by preserving its merits.

F&G-experiments have speeded up undoubtedly the process of a kind of unification of social sciences by employing some changes in types of preferences in the initial conditions of RCT, but Liberman-experiments and occasional findings of framing effects in behavioral economics remind that this could perhaps not be enough. MFS is a conjecture to integrate framing-effects into a comprehensive concept of a 'general' theory of action – even if a majority of cases can be explained actually by means of (wide) RCT alone. But a rigorous experimental test is still missing. What could be done, however, as a first next step seems to be obvious: combining the F&G-experiments with the Liberman-experiments and varying the easiest possibility to restrict opportunities for deliberation, namely time pressure to react in situations defined by different cues for frames with different types of actors.

15.10 Epilogue

Many have always viewed the MFS with skepticism. And there have been periods, partly up to date, when differences seemed to increase and when mutual rationalizations of the superiority of one's own approach have started to prevail once more (see Fehr and Gintis 2007: 60f., Gintis 2017: 161ff., Chapter 12). There are certainly good reasons for such resistances, as we know from Lakatos' work: As long as one has an alternative useful theory: formally precise, logical in substance, empirically proven, one won't change it only because single anomalies occur. This even applies in case that anomalies increase in number – as long as no better alternative is in sight and as long as one can hope that the problems can be solved by (more or less: marginal) changes in the established approach. And this is exactly what has been done for a long time: extension of RCT by introducing ever more additional motives and, finally, the withdrawal to a purely formal definition of RCT by, for example, certain axioms. But the rather flat waves of a slowly upcoming flood of hardly adjustable anomalies also from game theoretical experiments take their toll on even core assumptions of RCT. In addition, it is obvious that prominent representatives, who are well aware of all RCT's possibilities, themselves attempt to solve these problems (cf. Tutić 2015 on the approach of an integrative microfoundation from various convergences regarding developments within economy, cognitive (social) psychology, and sociology). It looks as though these developments have meanwhile also influenced some criticasters, at least there are indications that they have in the meantime been acknowledged as noteworthy developments in the microfoundation of the model of sociological explanation (Raub and Voss 2017: 29 f.). Not everybody has this open-mindedness and self-confidence to admit, that there is perhaps more

that has been established by now. I wish to thank Werner Raub especially, who, unlike others, was willing to do so.

References

Coleman, James S. 1986. "Social Theory, Social Research, and a Theory of Action." *American Journal of Sociology* 91:1309–1335.

Costa, Albert, Alice Foucart, Sayuri Hayakawa, Melina Aparici, Jose Apesteguia, Joy Heafner, and Boaz Keysar. 2014. "Your Morals Depend on Language." *PLoS ONE 9(4), e94842*. doi: 10.1371/journal.pone.0094842.

Diekmann, Andreas, and Thomas Voss. 2008. Die Bedeutung „sozialer" Motive für die Rational-Choice-Erklärung sozialer Normen. Pp. 83–7100 in: Andreas Diekmann, Klaus Eichner, P. Schmidt, and Thomas Voss (eds.), *Rational Choice: Theoretische Analysen und empirische Resultate. Festschrift für Karl-Dieter Opp zum 70. Geburtstag*. Wiesbaden: VS Verlag für Sozialwissenschaften.

Dufwenberg, Martin, Simon Gächter, and Heike Hennig-Schmidt. 2011. "The Framing of Games and the Psychology of Play." *Games and Economic Behavior* 73(2):459–748.

Ellingsen, Tore, Magnus Johannesson, Johanna Mollerstrom, and Sara Munkhammar. 2012. "Social Framing Effects: Preferences or Beliefs?" *Games and Economic Behavior* 76(1):117–130.

Engel, Christoph, and David G. Rand. 2014. "What Does 'Clean' Really Mean? The Implicit Framing of Decontextualized Experiments." *Economics Letters* 122(3):386–389.

Esser, Hartmut. 2017. "When Predictions Fail. Reactions of Rational Choice Theory and Behavioral Economics to the Unexpected Appearance of Framing Effects." Pp. 505–75026 in: Ben Jann, and Wojtek Przepiorka (eds.), *Social Dilemmas, Institutions and the Evolution of Cooperation*. Berlin and Boston: De Gruyter/Oldenbourg.

Esser, Hartmut. 2018. Sanktionen, „Reziprozität und die symbolische Konstruktion einer Sanktions-´Gemeinschaft´. Ein theoretischer Vergleich und empirischer Test von Rational-Choice-Theorie und dem Modell der Frame-Selektion anhand von Befunden und Daten aus der experimentellen Spieltheorie zur Erklärung der Bereitstellung von Kollektivgütern." *Zeitschrift für Soziologie* 37: 8–28.

Esser, Hartmut, and Clemens Kroneberg 2015. "An Integrative Theory of Action: The Model of Frame Selection." Pp. 63–85 in Edward J. Lawler, Shane R. Thye, and Jeongkoo Yoon (eds.), *Order on the Edge of Chaos. Social Psychology and the Problem of Social Order*. Cambridge: Cambridge University Press.

Fehr, Ernst, and Simon Gächter. 1999. Cooperation in Public Goods Experiments. Working Paper No. 10. Institute for Empirical Research in Economics. University of Zürich.

Fehr, Ernst, and Herbert Gintis. 2007. "Human Motivation and Social Cooperation: Experimental and Analytical Foundations." *Annual Review of Sociology* 33:43–64.

Gintis, Herbert. 2007. "A Framework for the Unification of the Behavioral Sciences." *Behavioral and Brain Sciences* 30:1–61.

Gintis, Herbert. 2017. *Individuality and Entanglement. The Moral and Material Basis of Social Life, Princeton NJ und Woodstock*. Oxfordshire: Princeton University Press.

Hermann, Benedikt, Christian Thöni, and Simon Gächter. 2008. "Antisocial Punishment Across Societies." *Science* 319: 1362–1367.

Keizer, Kees, Siegwart Lindenberg, and Linda Steg. 2013. "The Importance of Demonstratively Restoring Order." *PLoS ONE 8(6). e65137*, doi:10.1371/journal. pone.0065137.

Liberman, Varda, Steven M. Samuels, and Lee Ross. 2004. “The Name of the Game: Predictive Power of Reputations versus Situational Labels in Determining Prisoner’s Dilemma Game Moves.” *Personality and Social Psychology Bulletin* 30(9):1175–1185.

Lindenberg, Siegwart. 2015. “Social Rationality and Weak Solidarity. A Coevolutionary Approach to Social Order.” Pp. 43–62 in Edward J. Lawler, Shane R. Thye, and Jeongkoo Yoon (eds.), *Order on the Edge of Chaos. Social Psychology and the Problem of Social Order*. Cambridge: Cambridge University Press.

Rand, David G., Joshua D. Greene, and Martin A. Nowak. 2012. “Spontaneous Giving and Calculated Greed.” *Nature* 489:427–30.

Raub, Werner. and Thomas Voss. 1981. *Individuelles Handeln und gesellschaftliche Folgen, Darmstadt: Luchterhand*. Soziologische Texte, Neue Folge, Band 120.

Raub, Werner, and Thomas Voss. 2009: “Lob des Modellbaus.” Pp. 167–198 in Paul Hill, Frank Kalter, Johannes Kopp, Clemens Kroneberg, and Rainer Schnell (eds.), *Hartmut Essers Erklärende Soziologie. Kontroversen und Perspektiven*. Frankfurt and New York: Campus Verlag.

Raub, Werner, and Thomas Voss. 2017: “Micro-Macro-Models in Sociology: Antecendents of Coleman’s Diagram.” Pp. 11–36 in Ben Jann, and Wojtek Przepiorka (eds.), *Social Dilemmas, Institutions and the Evolution of Cooperation. Festschrift for Andreas Diekmann*. Berlin and Boston: De Gruyter/Oldenbourg.

Rubinstein, Ariel. 2013. “Response Time and Decision Making. An Experimental Study.” *Judgement and Decision Making* 8(5):540–551.

Schelling, Thomas S. 1960. *The Strategy of Conflict*. Cambridge, MA: Cambridge University Press.

Schütz, Alfred. 1971. „Das Wählen zwischen Handlungsentwürfen.“ Pp. 22–69 in Alfred Schütz, *Gesammelte Aufsätze, Band 2: Studien zur soziologischen Theorie*, Den Haag: Martinus Nijhoff.

Simon, Herbert A. 1983. *Reasons in Human Affairs*. Stanford, CA: Stanford University Press.

Tutić, Andreas. 2015. “Warum denn eigentlich nicht? Zur Axiomatisierung soziologischer Handlungstheorie.” *Zeitschrift für Soziologie* 44(2):83–98.

Tversky, Amos, and Daniel Kahneman. 1981. “The Framing of Decisions and the Psychology of Choice.” *Science, New Series* 211(4481):453–458.

Vanberg, Viktor J. 2002. “Rational Choice vs. Program-Based Behavior: Alternative Theoretical Approaches and their Relevance for the Study of Institutions.” *Rationality and Society* 14: 7–54.

Christoph Engel and Axel Ockenfels

16 Maverick: Experimentally Testing a Conjecture of the Antitrust Authorities

Abstract: Antitrust authorities all over the world are keen on the presence of a particularly aggressive competitor, a "maverick". Yet there is a lack of theoretical justification. One plausible determinant of acting as a maverick is behavioral: the maverick derives utility from acting competitively. We test this conjecture in the lab. In a pretest, we classify participants by their social value orientation. Individuals who are rivalistic in an allocation task indeed bid more aggressively in a laboratory oligopoly market. This disciplines incumbents. We conclude that the existence of rivalistic attitudes may justify antitrust policies that protect mavericks.

16.1 Introduction

One man's meat is another man's poison, as they say. Antitrust is a field of application. For those forming a cartel, or coordinating tacitly, collusion is a dilemma. Individually, each is best off if the others are faithful cartelists, while this one firm undercuts price, or exceeds the quota for that matter. However, if cartelists succeed to coordinate, this has negative external consequences for consumers. Antitrust authorities are therefore pleased to learn that one supplier in a market is particularly aggressive. The US Horizontal Merger Guidelines have coined the graphic term "maverick" for such firms. The Guidelines describe such firms as "*firms that are unusually disruptive and competitive influences in the market*".[1] The European Horizontal Merger Guidelines express the same concern.[2]

1 57 FR 41552, sec. 2.12 at note 19; the concern is upheld in the new, 2010, version of the guidelines, http://www.justice.gov/atr/public/guidelines/hmg-2010.pdf, sec. 2.1.5 and sec 7.1

2 OJ 2004 C 31/5, no. 20, no. 42.

Notes: Helpful comments by Uwe Cantner, Dominik Grafenhofer, Marco Kleine, Vincenz Frey and several anonymous referees, and discussion at the Rotterdam Department of Economics, are gratefully acknowledged. Ockenfels gratefully acknowledges financial support from the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) through the Research Unit "Design & Behavior" (FOR 1371) and under Germany's Excellence Strategy (EXC 2126/1–390838866), as well as from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (GA No 741409). This paper reflects only the authors' view and the funding agencies are not responsible for any use that may be made of the information it contains.

Christoph Engel, Max Planck Institute for Research on Collective Action, Bonn
Axel Ockenfels, Department of Economics, University of Cologne

https://doi.org/10.1515/9783110647495-016

In the next section, we review the case law and the (rather small) economic literature on maverick behavior. In this paper, we focus on one potential source of aggressive market behavior that has gotten short shrift. Market participants might bid aggressively because they hold particularly competitive preferences. They might derive utility from getting a higher payoff than their peers. In this sense, our study looks at macro-level implications of individual social preferences and thus builds on most of the literature, which asserts that such preferences exist in the field (see references in Ockenfels et al. 2015).

A preference-based explanation for aggressive market behavior, and its effect on the behavior of other market participants, would be hard to study in the field, if not impossible, though. This is why our study is conducted in a controlled laboratory environment, despite the inevitable wedge between our object of interest (the behavior of firms in a product market) and our object of study (the behavior of students in a laboratory market); we further discuss external validity in the concluding section.

Social preferences are assumed to be personality traits. Personality traits cannot be induced on the spot, but they can be measured. We proceed in two steps. In a first experiment, we classify participants by their social value orientation (Liebrand and McClintock 1988). We select those participants with the most rivalistic social value orientation to be entrants in the second, main experiment. For 10 periods entrants observe how two participants randomly selected from a pool with less extreme social value orientation choose quantities in a duopoly market. We investigate whether the behavior of incumbents, and market outcomes, differ according to the social value orientation of the entrant.

Our main hypothesis is supported with a proviso. Conditional on local market conditions, firms perform worse on average, and consumer welfare increases, if the market entrant is classified as rivalistic. Yet local conditions matter. In particular, rivalistic entrants do not make the market more competitive if competition was already fierce in the first place.

The remainder of the paper is organized as follows: section 2 defines our contribution to the legal and economic literature. Section 3 presents the design of the experiment and our hypotheses. Section 4 reports the results from the main experiment. Section 5 concludes with discussion.

16.2 Mavericks in practice and in economics

The concept of mavericks has led to a rather rich case law. In *United States vs. ALCOA*, government sued ALCOA for divestiture of the acquisition of Rome Cable Corporation. The Supreme Court held that the acquisition constituted monopolization, on the argument that “Rome was an aggressive competitor” (377 U.S. 271 [281] (1964)). Likewise, in *Mahle GmbH*, the Federal Trade Commission forced Mahle

GmbH to divest Metal Leve's United States piston business on the argument that, before the merger, Metal Leve was "an aggressive and innovative competitor" (62 Fed.Reg. 10,566 [10,567] (1997)). The Antitrust Division of the Department of Justice opposed the acquisition by Alcan Aluminium Corp. of Pechiney Rolled Products, LLC, since this would "remove a low cost, aggressive, and disruptive competitor in the North American brazing sheet market" (Case No. 1:03CV02012, para. 21 (2003)).[3] Likewise, the Federal Trade Commission opposed the proposed merger of *Staples, Inc.* with *Office Depot, Inc.*, on the argument that the merger would eliminate a "particularly aggressive competitor in a highly concentrated market" (Case No. 1:97CV00701, sec. IV A 2 (1997)). These decision are echoed by legal doctrine (Baker 2002; Kolasky 2002).

The European antitrust authorities have taken similar decisions. The European Commission cleared the merger of *T-Mobile Austria* with *tele.ring* only after the parties committed to selling major assets of *tele.ring* to an independent competitor. This undertaking was requested, although the new merged unit would not be the largest supplier in the Austrian market for the provision of mobile communication services to end customers since, before the merger, "for the last three years, tele.ring has played by far the most active role on the market in practising successfully a price aggressive strategy" (case M.3916, O.J. L 88/2007, 44, para. 10). Likewise the Commission cleared the merger of *Linde* with *BOC* only after both firms committed to selling a number of major supply contracts concerning helium. This removed the Commission's original concern that, otherwise, *Linde* would stop "compet[ing] aggressively to expand its position on this market" (case M.4141, IP/06/737 (2006)). An interesting case is *Euler Hermes/OEKB*. Through the merger, the new unit reaches a share between 45 and 55% on the Austrian market for delcredere insurance. The Commission nonetheless does not see reason for concern, one counter argument being that an independent new entrant *Atradius* "has assumed the role of a maverick by its aggressive pricing policy and its increase of sales" (case M.4990, para. 29, 2008).[4]

There is also empirical data suggesting that mavericks exist, and that they can substantially change market behavior. One study compares prices for retail gas in the otherwise comparable metropolitan areas of Ottawa and Vancouver. In both regions, tacit collusion would be equally feasible. Yet data from Internet price data collection sites show that, in the Ottawa region, prices are much more dispersed and volatile. This market outcome can be traced back to the presence of a maverick (Eckert and West 2004a, b). Maverick behavior has also been identified in the Australian mortgage market (Breunig and Menezes 2008). Another illustration is behavior in the Dutch spectrum auction in 2000 (Van Damme 2003, see also Klemperer 2004). There were five incumbents and five licenses for sale, but several potential entrants. As Van Damme (2003) emphasized, the Dutch telecom regulator "hinted at

3 http://www.justice.gov/atr/cases/f201300/201303.pdf.

4 http://ec.europa.eu/competition/mergers/cases/decisions/m4990_20080305_20310_de.pdf.

the desirability to favor newcomers to the market in the auction", and that "there are several reasons why a new entrant might be a more aggressive player on the market". However, all but one potential entrant (Versatel) actually partnered with an incumbent bidder, removing them from the auction market. One of the incumbents (Telfort) later, during the auction, accused Versatel of particularly aggressive bidding behaviors. As Van Damme (2003:285) reports: "Telfort claims that Versatel is bidding only to raise its rivals' costs or to get concessions from them." (Cramton and Ockenfels 2017 make a related point in the context of Germany's 4G auction.)

That said, there is a gap between the practice of dealing with mavericks in competition policy and the economics of mavericks in theory. Simple economic explanations of why some firms are more competitive than others would include that mavericks have lower costs, are incentivized by sales volumes, or control more capacities than their competitors. All this would imply that mavericks have a rather large market share. Yet, as Breunig and Menezes (2008) pointed out, competition authorities often stress that mavericks are, in fact, likely to be small firms (which seems to make it more plausible that personality traits of managers play a role in the phenomenon of mavericks). This might follow from pronounced switching cost, which forces entrants to be particularly aggressive (Farrell and Klemperer 2007), from more pronounced discounting of future earnings by firms in financial distress (Busse 2002), or from the fact that fixed cost is high in the industry (Scherer and Ross 1990). Yet another, underexplored source of aggressive behavior is behavioral. Some, but not all, decision makers like to be ahead, and dislike being behind. It is this source we are studying in this paper.

Our approach resonates with the New Zealand Merger Guidelines. In their section 7.2, the guidelines explicitly list "features associated with a maverick". Most features relate to a behavioral tendency to disrupt coordination and similar phenomena, including the first feature ("a history of aggressive, independent pricing behavior") and the last feature ("a history of independent behavior generally").[5] In the same spirit, Kwoka (1989) adds a firm specific degree of conjectural variation in quantity choices to a fully symmetric Cournot model.

In the US the focus on "maverick" firms has come under attack. Antitrust authorities have been urged to put less weight on the issue, mostly because there is so little theoretical foundation in economics.[6] However, in our view, the normative debate of the role of mavericks would benefit if it were to adopt a more adequate concept of competitive behavior. Individuals strongly differ with respect to social behavior, including their competitiveness, willingness to cooperate or collude, and

5 http://www.comcom.govt.nz/assets/Imported-from-old-site/BusinessCompetition/MergersAcquisitions/ClearanceProcessGuidelines/ContentFiles/Documents/Mergers-and-AcquisitionsGuidelines-2003.pdf, accessed 1 January 2014.

6 Personal communication by the chief economist of the German Cartel Authority, Konrad Ost.

ability to coordinate. In fact, individual heterogeneity in social and economic interaction is one of the most robust insights from behavioral economics and psychology (e.g. Camerer 2003). Thus, heterogeneity of social preferences may be one important missing link between antitrust practice and economic theory when it comes to understanding the presence of mavericks.[7]

There are many ways of modeling social preferences (for a survey see Cooper and Kagel 2016). Many models include a concern about relative, not only absolute payoff. Such models describe, for instance, inequity averse players (Fehr and Schmidt 1999; Bolton and Ockenfels 2000) or rivalistic players, who are willing to trade some absolute payoff against a sufficiently higher relative payoff (Fouraker and Siegel 1963: chapter 9; Bolton 1991; Frank 1984; Bazerman, Loewenstein, and White 1992; Messick and Thorngate 1967). These models resonate with an extended literature in social psychology on the "desire to win" (for a summary see Malhotra 2010). There is pronounced heterogeneity with respect to this desire (De Dreu and Boles 1998; Van Lange et al. 1997). The desire to win can lead to bidding more in an auction than the item is worth (Ku, Malhotra, and Murnighan 2005) and to engage in costly litigation rather than settling a case (Malhotra, Ku, and Murnighan 2008).

Rivalistic behavior is also sometimes characterized as status seeking (Frank 1985; Clark, Frijters, and Shields 2008) and backed by solid experimental evidence (Ball and Eckel 1998; Huberman, Loch, and Önçüler 2004; Charness, Masclet, and Villeval 2013) and evidence from the field (Solnick and Hemenway 1998; Ferrer-i-Carbonell 2005; Luttmer 2005; Boes, Staub, and Winkelmann 2010). The concept of status seeking has explicitly been extended to market behavior (Sobel 2009), entrepreneurial risk-taking (Clemens 2006) and managing a firm (Auriol and Renault 2008). Status seeking has been shown to affect behavior in experimental markets (Ball et al. 2001) and experimental supply chains (Loch and Wu 2008). In the field, status plays a strong role in motivating managers (Ockenfels, Sliwka, and Werner 2014; Grund and Martin 2017).

The only experimental study of "maverick" behavior we are aware of has been conducted by Li and Plott (2009). The paper studies which interventions can break tacit collusion in a laboratory market with 8 participants who hold exogenously given, different valuations for 8 items. The first part of their experiment continues until the group colludes perfectly. One of the interventions, which the authors relate to the anti-trust concept of a maverick, consists of confidentially changing the valuations of 2 items for the duration of 2 periods. As desired, participants with higher valuations, who have been induced to bid more aggressively, start bidding for the item in question. Some other participants retaliate, which leads to a price

7 Of course, other areas of industrial organization have already been substantially influenced by behavioral research; see, e.g., Engel (2007) for the insights from experimental economics for the determinants of tacit collusion.

war. Yet after a while, collusion is again established (Li and Plott 2009: 444). Our approach complements their study in various important ways. We mention two points here. First, we study the effects of aggressive quantity choices *resulting from personality*. That is, our study does not induce aggressive behavior by confidentially changing monetary incentives, but rather focuses on the potential of naturally occurring heterogeneity in social motivation to capture maverick behavior. Indeed, because in our context all payoff functions and market conditions are identical and common knowledge across subjects, heterogeneous individual traits are the only possible cause for treatment effects in our experiment. Second, we investigate the effect of "maverick" behavior in markets that, endogenously, have produced *different degrees of competition*. As we will see, our variables of interest matter: market outcomes can be related to natural psychological traits of traders, and the impact of maverick behavior interacts with idiosyncratically evolved market competitiveness.

Our paper also makes a contribution to the experimental literature on social dilemmas. To the best of our knowledge, no experiment has tried to explain outcomes in oligopoly markets with the social preferences of participants (cf. the theory paper by İriş and Santos-Pinto 2014). This is surprising given competition can be modelled as a dilemma, and choices in dilemma games are routinely rationalized with the social preferences of participants (for a survey see Chaudhuri 2011). We do not only derive hypotheses from participants' social preferences, but even build our treatment manipulation on randomly composing markets conditional on participants' social preferences.

16.3 Design of the experiment and hypotheses

In order to test the effect of heterogeneous preferences on competition we first classify participants according to their social value orientation in a pre-test, using the standard procedure introduced by Liebrand and McClintock (1988). This test has participants repeatedly choose between two different allocations of a sum to be distributed between an anonymous partner and themselves. They are, for instance, asked whether they prefer 354 units for themselves and an anonymous counterpart over 397 units for themselves and 304 units for the counterpart. Aggregating over all 32 incentivized choices, for each individual one defines a score, which is customarily called the "ringdegree" since the measure can be represented on a circle. Participants with a score of 0 only care about their own payoff. Participants with a positive score are willing to give up some payoff for themselves for the sake of giving their anonymous partner a higher payoff. Such participants are averse against advantageous inequity, consistent with Fehr and Schmidt (1999), Bolton and Ockenfels (2000). We are particularly interested in participants with a negative score. They are willing to give up some payoff for themselves in the interest of increasing the payoff

difference between themselves and their partner. These participants are rivalistic. They hold a positive willingness to pay for improving their status.

In the main experiment, we form fixed markets of three suppliers to interact in a fully symmetric Cournot market over 20 rounds. In the first 10 rounds, only two suppliers, the incumbents, are active. Every round, the passive supplier, the entrant, is informed about price and total quantity. This participant only enters the market in round 11. This procedure allows the entrant to observe the market before entering, which seems reasonable for any potential entrant. The social value orientation of the entrant is our treatment variable. We have rivalistic entrants, selfish entrants, and entrants who are averse against advantageous inequity. This design reflects the fact that social value orientation, as a personality trait, is not open to ad hoc manipulation. The trait can only be measured, and participants can be matched by the trait. While we are not aware of experiments that have used this approach for social value orientation, it is, for instance, common if one uses gender, age or race as treatment variables (for references in dictator game experiments see, for example, Engel 2011).

We emphasize that, with the design of the experiment, we do not identify the effect of the presence or absence of a maverick on competition. What we measure is the effect of a change in the structure of the market through the market entry of a maverick. We are thus testing a dynamic, not a static effect (on this distinction see Engel 2016), akin to the distinction between stocks and flows. We have chosen this research question for reasons of external validity. Antitrust, and merger control in particular, have been primarily interested in preserving the competition enhancement resulting from such market entry.

The social value orientation test is run a couple of days before the market experiment. Participants are invited on the understanding that a second experiment is to follow, but are not informed about the nature of the second experiment. To make matching in the main experiment possible, but preserve anonymity, we use the following procedure: at the end of the pre-test, participants themselves generate an identification code. Participants write this code on a card, put this card into an envelope, seal the envelope and write their name on it. The closed envelopes go to the lab manager. The manager opens them and writes a list that matches names and codes. The experimenter prepares a list with groups to be invited for the main experiment. In this list, participants are only identified by their code. The lab manager does not learn any choices participants have made, neither in the pre-test nor, later, in the main experiment. The lab manager only knows who shall be invited for which session. The experimenter never sees the list that matches codes and names. At the outset of the main experiment, participants identify themselves on the computer screen by their code. The program checks whether the invited participants are present.

Participants are completely informed about this procedure. They also know that the experiment has two parts, and may therefore infer that information from

the pre-test is used for inviting participants to one of the sessions of the main experiment. Yet participants neither know the nature of the main experiment, nor which information is used for matching (we run a battery of further personality tests the results of which are of no relevance for the main experiment; their only purpose is making it difficult for participants to infer which personality trait is used for matching).[8] In particular, subjects are neither informed about behavior in the first experiment nor about social value scores of other participants; in the field, too, other firms usually only observe their competitors' behavior, not their preferences or decision making process.

In the main experiment, participants interact in fixed groups of three. The main experiment has two parts.[9] At the outset, participants only receive instructions for the first part. They are informed that more parts are to follow, and that new instructions will be distributed for the continuation. The first part of the main experiment has 10 rounds. In this part of the experiment, two incumbents of each group have the active role. The entrant has the passive role. Incumbents are not told that the third participant will later enter the market. This design feature is meant to capture the situation when maverick behavior is most important for antitrust: an outsider observes whether aggressive market behavior is likely to be profitable. (We note, however, that being worried about entry could have led to stronger competition and thus reduce the effect of a maverick entrant.) Incumbents compete in a Cournot market where the profit of incumbent i in period t is given by (16.1).

$$\pi_{it} = \left(100 - q_{it} - q_{jt}\right) q_{it} \tag{16.1}$$

We thus assume demand to be linear and normalize cost to zero. After each period, incumbents learn the resulting price and their individual profit. Entrants learn total quantity supplied and the price. After the end of period 10 there is a (surprise) restart of the market. Now entrants become active as well, so that the profit function changes to (16.2).

$$\pi_{it} = \left(100 - q_{it} - q_{jt} - q_{kt}\right) q_{it} \tag{16.2}$$

The second part of the experiment also lasts 10 periods.

8 In the pre-test, we had the following sequence of tests: social value orientation; risk preferences (Holt/Laury); belief elicitation on 4 problems from the test for social value orientation; Big5 personality inventory (short 10 item version); 4 unincentivized questions about trust taken from the German socio-economic panel; basic demographic information.

9 Plus a third part meant to test a theoretical prediction that buyouts of the entrant will not occur, which has been confirmed in our data. However, because this is only of secondary importance for our results, we decided to drop this part altogether. We refer interested readers to the working paper version of this article for more details.

Based on the results of the pre-test, three groups of participants are selected to have the entrant role in the main experiment: Those 9 participants with the most negative social value orientation score have the entrant role in the *Negative* treatment. These participants are rivalistic. We form two different comparison groups: 11 participants with a social value orientation score of zero have the entrant role in the *Zero* treatment. These participants are selfish. Those 11 participants with the highest positive social value orientation score have the entrant role in the *Positive* treatment. The remaining participants are randomly assigned to have the incumbent role in either treatment. Three of them have a mildly negative social value orientation score. 16 of them are selfish. 40 have a mildly positive social value orientation score.[10]

We have 9 groups (27 participants) in the *Negative* treatment, and 11 groups (33 participants) in the remaining two treatments. Participants are invited using the software ORSEE (Greiner 2004). 52% of participants are female. Average age is 25.45 years.[11] Participants, most of whom are students, hold various majors. The experiment is programmed using the software zTree (Fischbacher 2007). It is run in the Bonn EconLab. In the pre-test, participants on average earn 13.20€ (16.05$ on the days of the experiment). In the main experiment, they on average earn 9.36€.[12]

We can straightforwardly compute our null hypothesis under the standard assumption that all suppliers maximize their individual payoffs. There is a unique subgame-perfect equilibrium strategy for each phase of the experiment,[13] conditional on the number of suppliers in the market, which is given by $q_i = \frac{100}{n+1}$, where n is the number of suppliers. Plugging in the respective market size, we get our null hypothesis

> **H_0:** Participants' social preferences for competitiveness do not affect market outcomes; only market size matters.

For our alternative hypothesis, assume that there is some heterogeneity of preferences. In particular, assume that the entrant is a maverick, competing more aggressively than

10 The fact that three participants with a negative social value orientation score are incumbents results from a mistake of the lab manager. Since the lab manager did not know their social value orientation scores, these participants were randomly assigned to one of the groups. For five incumbents we do not know the social value orientation score. These subjects replaced invited participants who did not show up.

11 From the five replacement subjects, we do not have demographic information since the demographic questionnaire was part of the first experimental battery.

12 The tasks participants face in both parts of the experiment are unrelated, so that the difference in earnings across parts is not meaningful.

13 This is because each base game has a unique equilibrium. In fact, if at the beginning of the first phase, subjects had common knowledge about all aspects of the subsequent phase of the experiment, the subgame perfect equilibrium of the whole game could be computed, would also be unique and correspond to the equilibrium in each phase of the experiment.

standard theory would predict. Specifically, assume that the entrant not only cares about absolute profit but also about earning more than the competitors, and that this is common knowledge.[14] Then, like commitment power favoring the Stackelberg leader, the rivalistic supplier sells a larger quantity than in a standard analysis of the Cournot market, and the incumbents – if only interested in own gains – sell a smaller quantity. Total quantity and thereby consumer welfare is larger than if all suppliers hold standard preferences.[15] This leads to

> **H_1:** If the entrant is rivalistic, she sells higher quantities and the market outcome is more competitive.

We mention that we can derive the same hypothesis if we allow incumbents to be rivalistic, too, as long as they are less rivalistic than the entrant (see Appendix I).

16.4 Experiment results

Figure 16.1 informs about the distribution of social value orientation in our sample. We have 12 (13.64%) rivalistic, 27 (30.68%) selfish, and 49 (55.68%) participants with a more or less pronounced positive social value orientation.[16] Figure 16.1 also shows our matching. Participants at the lower end of the distribution are entrants in the *Negative* treatment. These are the subjects with the supposedly most competitive behavior in oligopoly markets, and they are thus the focus of our study on the impact of mavericks. Participants at the upper end of the distribution are entrants in the *Positive* treatment. 11 participants with a social value orientation score of zero are entrants in the *Zero* treatment. The remaining participants are randomly

14 This is a common assumption not only in large parts of the social preferences literature, but also in the economics literature that does not address social preferences. The assumption simplifies theoretical derivations, although it seems incorrect in most applications. However, in our setting any rivalistic motivation leads to more aggressive bidding, regardless of the extent to which competitors are (believed to be) rivalistic. In this sense, the general insight that rivalry leads to larger quantities is robust.

15 We focus on *consumer* welfare for two reasons. Enhancing consumer welfare is the primary stated goal of antitrust policy (Crandall and Winston 2003). Moreover we model mavericks as agents holding social preferences, so that the definition of supply side welfare is not obvious. By focusing on the opposite market side, we are able to bracket this debate in normative economic theory.

16 Social value orientation scores range from – 56.23 (strongly rivalistic) to 74.55 (strongly averse to advantageous inequity). If a participant chooses the allocation that gives her a higher payoff on all 32 problems, her score is 0. A participant with a score of 45 always chooses the equal split. A participant with a score of 90 is perfectly altruistic. A participant with a score of – 45 is willing to give up 1 unit of her absolute profit to increase the payoff gap between herself an her random partner by 1 unit. For the procedure for aggregating the 32 choices see Liebrand and McClintock (1988).

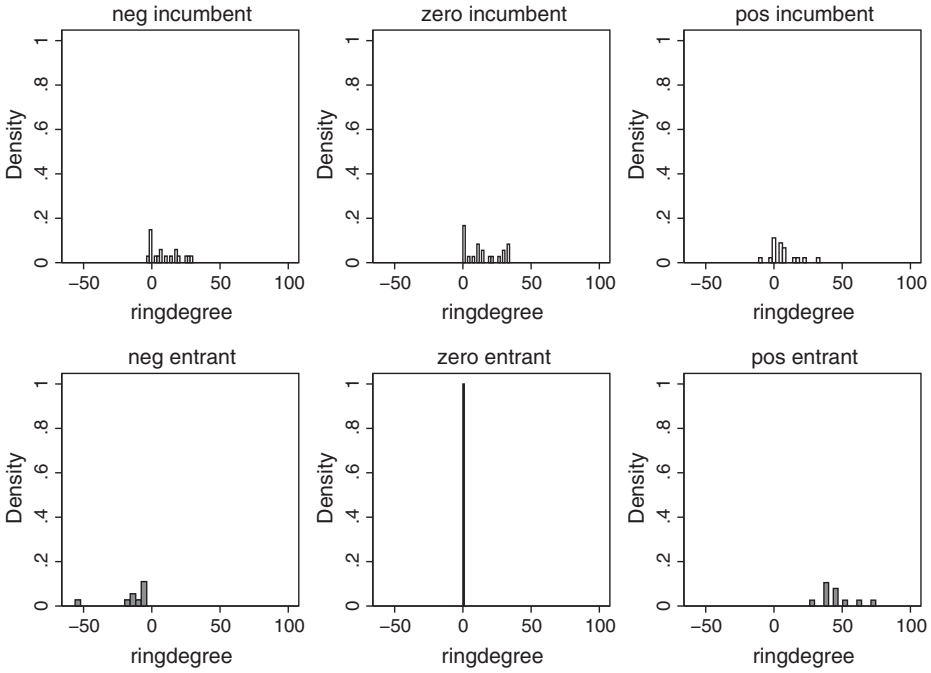


Figure 16.1: Social value orientation per treatment and role.

assigned to being incumbents in either treatment. To make sure that the 16 selfish incumbents are equally distributed across treatments, randomization is separate for participants with a social value orientation score of 0, and for the remaining incumbents.

As Figure 16.2 shows, overall quantity choices are fairly close to the standard Cournot predictions. In duopoly markets, average quantity is close to 33. In triopoly markets, it is close to 25. We thus provisionally support our null hypothesis $\mathbf{H_0}$. Looking at average quantities only, social value orientation is not a plausible candidate for identifying maverick behavior. As suggested by Figure 16.2 and Table 16.1, if we work with averages, we do not find treatment effects, neither non-parametrically nor parametrically.[17]

This also holds if we confine the analysis to the last period before and the first period after entry. Actually, descriptively in the *Negative* and in the *Positive* treatments, entrants on average even sell less than incumbents and consequently make a lower profit.

17 For non-parametric estimation, we use a Mann-Whitney test, for parametric estimation the regression as specified in Table 16.2, but of course without controlling for the average quantity in periods 1–10.

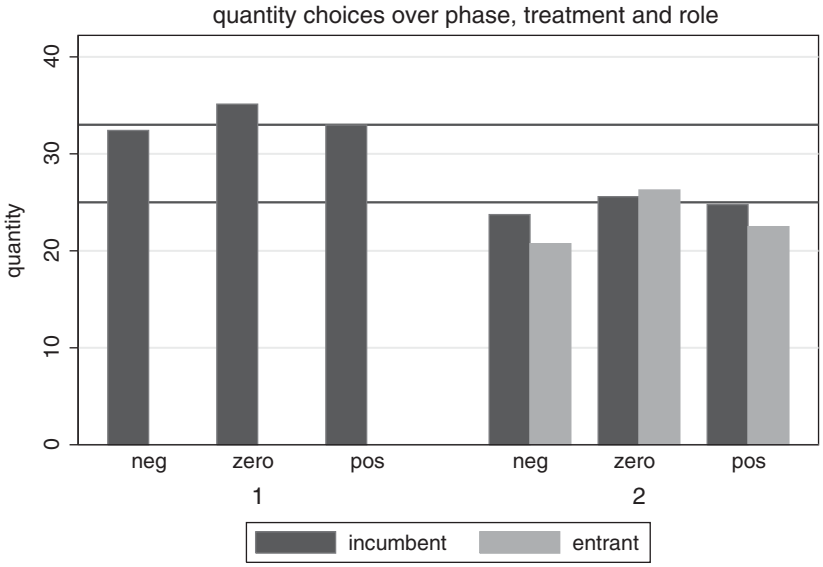


Figure 16.2: Aggregate quantity choices.

Table 16.1: Descriptive statistics.

	Phase 1			Phase 2		
	neg	**zero**	**pos**	**neg**	**zero**	**pos**
incumbent	32.411	35.123	32.973	23.733	25.582	24.773
	(7.564)	(13.905)	(7.641)	(6.451)	(13.794)	(8.732)
entrant				20.733	26.273	22.491
				(8.727)	(18.063)	(6.452)

Note: standard deviations in parenthesis

Yet, as Figure 16.3 illustrates, aggregates per treatment conceal a more complex story. In this figure, each marker is the mean quantity set by the two incumbents or the entrant in one group. There is quite some variation that is hidden by looking at averages only. In phase 1 of the Cournot market, quantity choices have mean 33.57, but standard deviation 10.34. Quantity choices in the second phase of the experiment heavily depend on experiences from the first phase. Independent of treatment, what the group has experienced while the market was a duopoly is a strong predictor of quantity choices after the entrance of the new competitor. Suppliers only adjust quantities to reflect greater competition: the trend line is close to 75% of the average quantity in the first 10 periods (which would be the quantity ratio of a triopoly compared to a duopoly, as predicted by standard theory). As the distribution of hollow (incumbents) versus solid markers (entrants) shows, market history matters for old

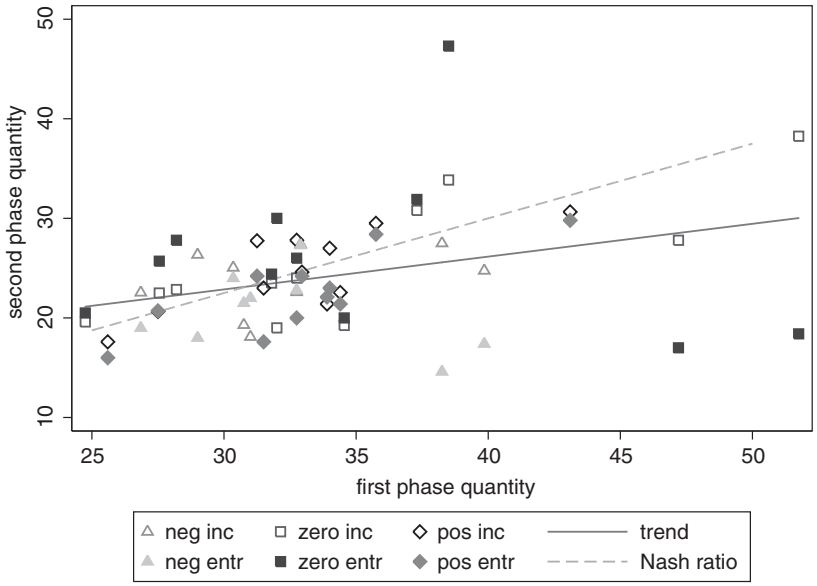


Figure 16.3: Dependence on local conditions.
Notes: x-axis: mean quantity sold by the two members of the duopoly, in periods 1–10
y-axis: mean quantity sold in periods 11–20
separately for incumbents (hollow markers) and for entrants (solid markers)
trend: linear prediction
Nash ratio: 3/4 of first phase quantity

and new market participants. We note that this history effect is in line with the only other experiment we are aware of that tests market entry (Goppelsroeder 2009). Overall, we can conclude that while there is a lot of idiosyncrasy regarding market competitiveness, Nash equilibrium goes a long way to predict *average* quantities and *average* differences of competitive pressure in our duopoly and triopoly settings.

The visual impression that local market competitiveness in periods 1–10 matters is supported by statistical analysis (see Table 16.2).[18]

18 We revert to regression analysis since we want to show that choices in periods 11–20 are explained by the average quantity this group had chosen before the third supplier enters the market. We have data from choices, nested in individuals, nested in groups. Dependence within individuals is captured by the random effect. The additional source of dependence at the group level is captured by clustering standard errors at this level. The fact that the Hausman test does not turn out significant shows that we are justified in preferring the more efficient random effects model over a model with individual fixed effects. The coefficient of the average quantity in phase 1 is smaller than 0.75 since the model has a constant. If we estimate the same model (as a population averaged regression) without a constant, the coefficient comes very close to the theoretical expectation and is 0.714.

Table 16.2: Effect of local conditions.

average quantity in periods 1–10	**0.414*** (0.110)**
Cons	10.367** (3.373)
N	930
p model	0.0002
R^2 within	0
R^2 between	0.1165
R^2 overall	0.0477

Notes: dependent variable: quantity, data from periods 11–20 random effects, robust standard errors clustered at the group level Hausman test insignificant on mirror model with period as additional regressor (to enable fixed effects estimation) standard errors in parenthesis * = p < .05

This gives us:

Result 1: If a new competitor enters a repeated Cournot duopoly market, higher pre-entry quantity is associated with higher post-entry quantity.

Knowing that local market conditions matter, we revisit the effects of our manipulation in Table 16.3.[19] The constant of the regression in Model 1 predicts the amount a firm would sell in the *Positive* treatment if the average amount sold in this group in the first 10 periods had been 0. Of course, as Figure 16.3 shows, in the experiment there has been no such market. The regression generalizes to the population of Cournot duopolies observed by entrants. If we plug in the average amount sold in the first 10 periods from all 11 markets where the entrant has a positive social value orientation score (32.973, Table 16.1), the regression predicts that, in the *Positive* treatment, firms on average sell 24.02 units,[20] which comes pretty close to the Nash quantity of 25 units.

From the significant positive main effects of treatments *Negative* and *Zero* we learn that, overall, the market is more competitive if the entrant is rivalistic or selfish, compared with a market where the entrant has a preference to avoid payoff differences. Yet this treatment effect is indeed conditional on the competitiveness before market entry. The significant negative interactions show that the translation effect is most pronounced if the entrant has a positive social value orientation

19 The fact that "overall" all models seem to explain little variance is an artefact of the fact that, by their design, these models only explain between, not within variance.

20 .638 + 32.973 * .709 = 24.02.

Table 16.3: Treatment effects conditional on local conditions.

	periods 11–20 all participants			period 11 entrants only	
	model 1	model 2	model 3	model 4	model 5
neg	19.427***	12.313^{+}	67.446**	102.635*	80.184*
	(4.674)	(6.851)	(24.389)	(46.349)	(29.398)
zero	12.664*	2.073	45.375*	34.287	52.523*
	(5.056)	(5.481)	(18.499)	(25.464)	(22.111)
entrant		−3.993	−3.993		
		(3.584)	(3.589)		
*neg**entrant		21.343^{+}	21.343^{+}		
		(12.874)	(12.888)		
*zero**entrant		31.775**	31.775**		
		(9.249)	(9.259)		
average quantity in period 10				1.044	
				(.730)	
average quantity in phase 1	.709***	.692***	1.990**		1.430*
	(.099)	(.114)	(.574)		(.572)
entrant*average quantity in phase 1		.052	.052		
		(.100)	(.100)		
*neg**average quantity in period 10				−3.051*	
				(1.401)	
*zero**average quantity in period 10				−1.029	
				(.767)	
*neg**average quantity in phase 1	−.627***	−.400*	−2.039**		−2.383*
	(.135)	(.197)	(.749)		(.894)
*zero**average quantity in phase 1	−.353*	−.078	−1.377*		−1.555*
	(.162)	(.168)	(.587)		(.652)
entrant**neg**average quantity in phase 1		−.680^{+}	−.680^{+}		
		(.383)	(.384)		
entrant**zero**average quantity in phase 1		−.823**	−.823**		
		(.270)	(.271)		
entrant SVO			1.061*		
			(.436)		
average quantity in phase 1*entrant SVO			−.032*		
			(.014)		

Table 16.3 (continued)

	periods 11–20 all participants			period 11 entrants only	
Cons	.638	1.969	−41.333*	−13.755	−27.069
	(3.314)	(3.844)	(18.080)	(23.815)	(19.017)
N	930	930	930	31	31
p model	<.0001	<.0001	<.0001	.420	.1633
R^2 within	0	0	0		
R^2 between	.1512	.2365	.2494		
R^2 overall	.0619	.0968	.1021	.0056	.1092

Notes: regression equations for all models in Appendix II dependent variable: quantity models 1–3: data from periods 11–20, models 4–5: data from period 11 models 1–3: data from incumbents and entrants, models 4–5: data from entrants only models 1–3: random effects, robust standard errors clustered at the group level Hausman test insignificant on mirror models with period as additional regressor (to enable fixed effects estimation) SVO: social value orientation, i.e. score from ring measure test treatment: reference category: *positive* standard errors in parenthesis *** p < .001, ** p < .01, * = p < .05, $^{+}$ p < .1

score. The more the market was competitive pre-entry, the less it becomes even more competitive through the entry of a new competitor with rivalistic or selfish preferences. In fact, the pro-competitive effect of the entrant holding rivalistic preferences only plays itself out if the average quantity pre-entry was at or below 31 units[21]; recall that the Nash quantity for the duopoly is 33 units. Likewise, if the entrant is selfish, entry only has a pro-competitive effect if the average quantity pre-entry was at or below 36 units.[22] Yet in both treatments, the pro-competitive effect of entry is pronounced if the duopoly was perfectly collusive. The model predicts that quantity is 3.752 units higher if a rivalistic firm enters a collusive market, and 3.839 units higher if a selfish firm enters.[23]

Model 2 splits the analysis by entrants and incumbents. The picture nicely clears if, in model 3, we additionally control for the precise social value orientation score of the entrant, and how it interacts with the competitiveness of the market before she enters. The following discussion focuses on this model. The implications are easiest to see in the marginal effect of the *Negative* and *Zero* treatments that are reported in Figure 16.4. If we find a significantly positive effect of treatment, the

21 19.427/.627 = 30.984.

22 12.664/.353 = 35.875.

23 19.427−25*.627 = 3.752; 12.664−25*.353 = 3.839.

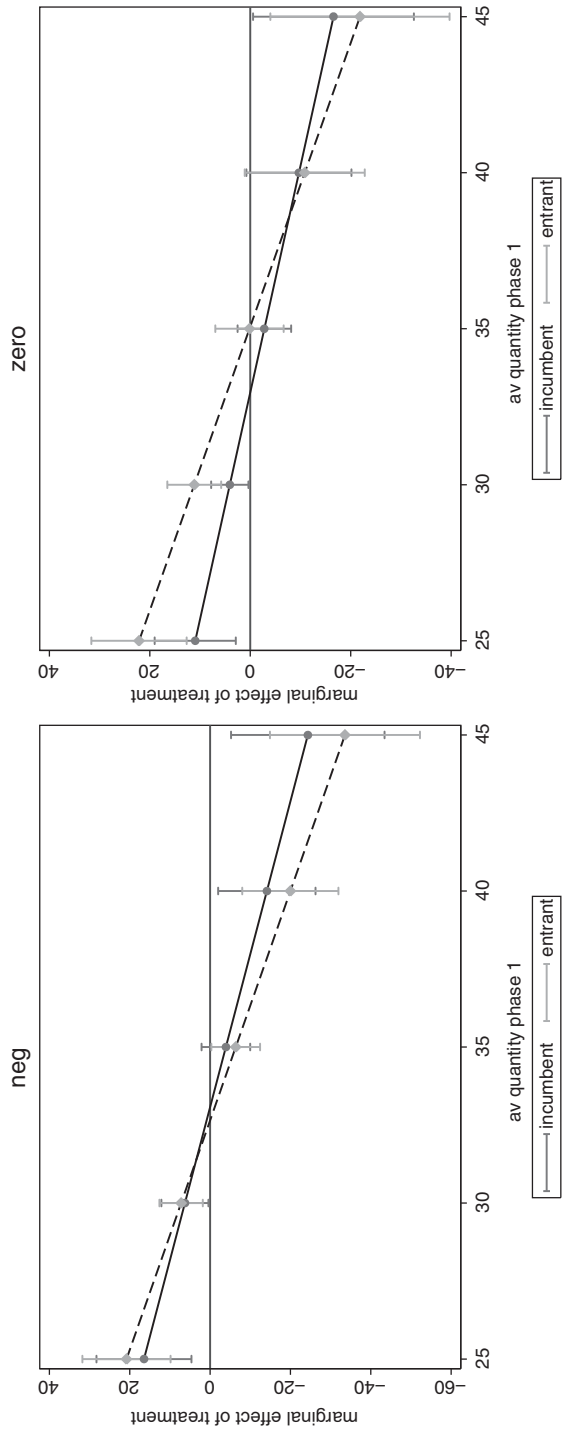


Figure 16.4: Marginal effect of treatment, conditional on role and competitiveness.
Note: marginal effects from model 3 of Table 16.3

rivalistic or selfish personality of the entrant has a pro-competitive effect. This holds true for both treatments and roles, but only if, pre-entry, the market was collusive.[24] With this qualification, we reject our null hypothesis $\mathbf{H_0}$ and infer that the alternative hypothesis $\mathbf{H_1}$ captures the data. We also note that the asymmetric response of selfish (*Zero*) and rivalistic (*Negative*) entrants to their observations from the first 10 periods is well in line with their playing best responses, assuming that incumbents will only adjust to the fact that one more supplier enters the market (but not reach equilibrium choices themselves). In the Appendix I we show this formally.[25]

To see whether the social preferences of entrants are indeed critical, we consider period 11 in isolation, i.e. the first period after entry. Overall, and if we confine the analysis to incumbents, we do not find any treatment effects, even if we interact treatment with the average quantity chosen in the respective group in period 10 (i.e. directly before entry), or during all of periods 1–10. But we do see a strong effect of the *Negative* treatment if we separately analyze choices of entrants (Model 4 of Table 16.3). We also see an effect of the *Zero* treatment if we replace average choices in period 10 with average choices in periods 1–10 (Model 5 of Table 16.3). Recall that incumbents had no information about the criterion for selecting entrants. Models 4 and 5 not only show that our manipulation worked. Together with Models 1–3 we also see how a maverick changes the market: immediately after entry, she behaves according to her preferences; in later periods, incumbents react to this experience.

Thus far our data suggest that a rivalistic and a selfish entrant have pretty much the same effect on competitiveness. To see whether this is indeed true, we use the following approach: individually for each incumbent we regress quantities sold in the first phase on time. This procedure gives us for each individual incumbent the trend, had there not been entry. From these regressions, for each individual we derive an out of sample prediction for the remaining 10 periods. We adjust the predicted quantity to the market entry of one more supplier by multiplying it by the theoretically predicted ratio of ¾ (see above). Note that the prediction is flat if, pre entry, the market had already reached equilibrium. However, inspecting the raw data, it seems that most duopoly markets had not yet stabilized. Only 17 of 62 incumbents did not change the quantity over periods 6–10.

Figure 16.5 shows the difference, per treatment and period, between the mean actual and predicted quantity. In the *Positive* treatment, actual quantities are much higher than the prediction. In the *Zero* treatment, actual quantities exhibit more variance, but have about the same level as the prediction. By contrast in the *Negative* treatment, and only in this treatment, for all periods but the final actual

24 The marginal effects of Figure 16.4 also explain the seemingly contradictory descriptive finding that, in the *Negative* treatment, entrants on average choose smaller quantities than incumbents, Table 16.1: entrants only bid more than incumbents if the market had been collusive.

25 We are grateful to an anonymous referee for suggesting this approach.

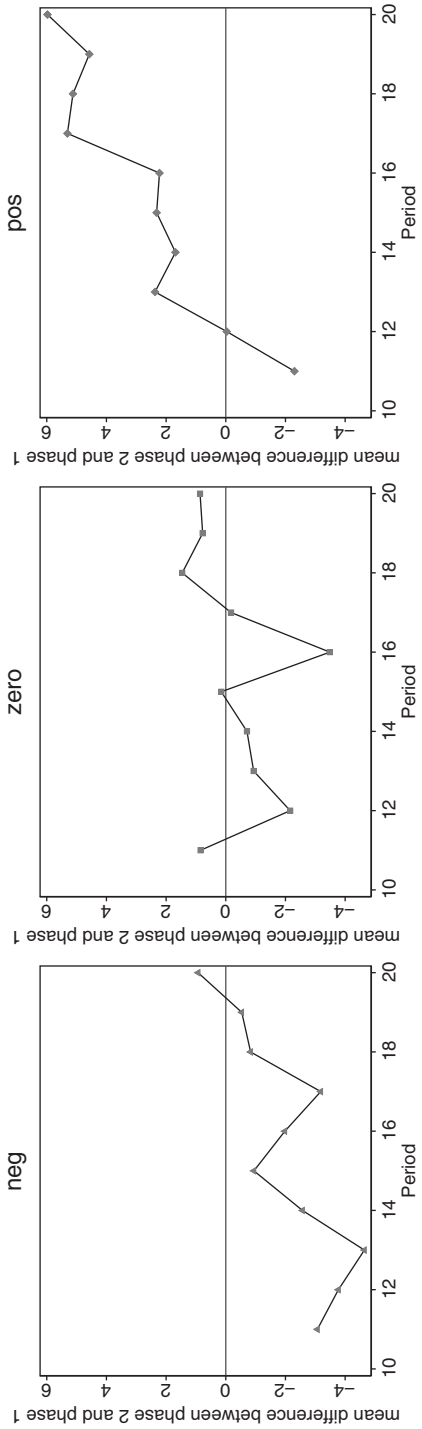


Figure 16.5: Effect of entry on choices of incumbents.
Notes: periods 11–20 only dependent variable: difference, per treatment and period, between the mean predicted and actual quantity

quantities are below the predicted trend.[26] We conclude that, depending on the social preferences of the entrant, incumbents come under additional competitive pressure and react by reducing the quantity they sell, as predicted by our model.

Overall, this gives us:

Result 2: Conditional on pre-entry local market competitiveness, a Cournot market is more competitive if the entrant is rivalistic.

In the final step, we want to understand in which ways rivalistic entrants discipline incumbents. To that end we take a closer look at dynamics in the *Negative* treatment. The dependent variable is changes in incumbents' choices from one period to the next.

Model 1 of Table 16.4 shows that incumbents, on average, reduce their own contributions in reaction to high contributions of the entrant ($p = .088$), as predicted by our theory. The weakly significant interaction effect ($p = .099$) indicates that the effect is the more pronounced the more the market was collusive before the third supplier entered. Model 2 and the marginal effects reported in Figure 16.6 show that the effect requires some degree of discord among the incumbents though.[27] If the standard deviation of quantity choices in this group and every period of phase 1 was low on average (range [2.828, 7.778]), incumbents do not significantly reduce their quantity in reaction to a high quantity sold by the entrant. This suggests that a duopoly that has successfully established a common norm of behavior is more resilient to attempts of a maverick to break up collusion; indeed, previous research has shown that homogeneous cooperation across agents is less vulnerable to being destabilized (e.g. Brosig, Weimann, and Ockenfels 2003).

16.5 Conclusion

Antitrust authorities are not only concerned with market power. They are also attentive to firm-specific heterogeneity in market behavior. They are particularly pleased if

26 The visual impression is supported by statistical analysis. If we regress the difference between the actual quantity and the out of sample prediction on treatment, and choose the *Negative* treatment as reference category, the constant informs us about the treatment effect for this treatment. If we use all 10 periods of the second phase, the constant is − 1.641, but not significantly different from zero ($p = .151$). If we repeat the analysis for periods 11–19, however, the constant is − 2.392, $p = .002$, which supports our claim. In neither regression, the net effect of constant + treatment *Zero* is significantly different from zero ($p = .9016$ in the first and $p = .8519$ in the second regression). We do, however, acknowledge that the treatment effect diminishes over time. If we repeat the regression, now interact treatment with period, and subsequently test the net effect of the constant + period, the result is significantly different from zero for periods 11–16 only. The additional regressions are available from the authors upon request.

27 Further controlling for the mean quantity sold individually by each incumbent, or replacing the standard deviation with this measure, does not yield significant effects.

Table 16.4: Reaction of incumbents to quantity choices of entrants in *neg* treatment.

	model 1	model 2
quantity sold by entrant in t-1	−3.094^{+} (1.590)	9.062^{+} (4.286)
quantity sold by entrant in t-1*average quantity in phase 1	.095^{+} (.051)	−.302^{+} (.136)
quantity sold by entrant in t-1*standard deviation of average quantity in phase 1		−2.237* (.774)
quantity sold by entrant in t-1*average quantity in phase 1*standard deviation of average quantity in phase 1		.073* (.025)
cons	1.559 (1.489)	−.753 (1.927)
N	162	162
R^2 within	.0749	.1117
R^2 between	.0847	.2194
R^2 overall	.0131	.0172

Notes: dv: quantity(t) – quantity(t-1) of incumbents data from *neg* treatment individual fixed effects, since Hausman test turns out significant robust standard errors, clustered for groups, in parenthesis *** p < .001, ** p < .01, * = p < .05, $^{+}$ p < .1

they identify especially aggressive firms. In this paper we experimentally investigate a cause for such "maverick" behavior that transcends pecuniary incentives: an individual may derive utility from relative, not only from absolute payoff.

In our experiment, we do indeed find that market entry by a participant with a particularly rivalistic attitude makes the market more competitive, improving consumer welfare and hampering incumbents' profits. Yet this result only holds conditional on the level of competition pre-entry. The entry of a "maverick" is socially most beneficial when it is most needed, i.e. when the market was collusive. This suggests that mavericks can play an important role for entertaining competitive markets, and so competition authorities may be indeed well-advised to appreciate this role in their policies.[28]

We of course do not claim a one to one mapping of the behavior of students in the lab (which we test) to the behavior of firms in markets. Firms are highly aggregate

28 The fact that we do not have even stronger findings might also result from the composition of our sample. In line with previous experimental results (Liebrand and McClintock 1988), only a minority of our participants is willing to give up some income for increasing the distance in payoff to their favor. With one exception, even those who do are only mildly rivalistic.

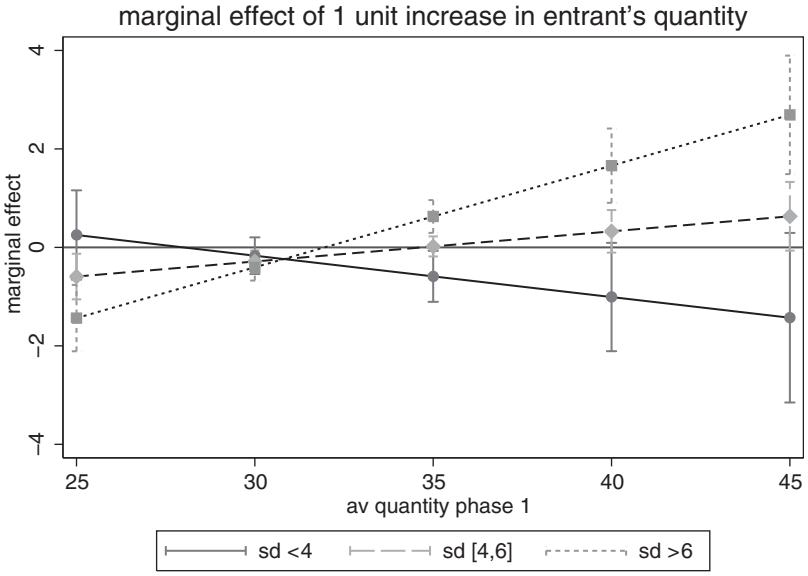


Figure 16.6: Reaction of incumbents to quantity choices of entrants in *neg* treatment.
Notes: marginal effects of 1 unit lagged increase in entrant's quantity on change in incumbents' quantity from model 2 of Table 16.4

corporate actors (for a survey of the experimental research specifically addressing such actors see Engel 2010), and decision making is rarely individual but rather based on some aggregation of team preferences; suppliers in a real market of three do not interact anonymously and underlying preferences of both, incumbents and mavericks may be subject to selection effects; and markets are differently organized and structured – to name only some obvious simplifications. But in line with a rich literature on experimental oligopoly markets (see the meta-study by Engel 2015) we believe that such evidence provides a useful starting point for analyzing the behavior of firms. Eventually, individuals decide for firms. It is therefore not unlikely that behavioral traits of these individuals carry over to the behavior of the firms for whom they act. Managers are not only selected for their competence and connections, but also for their personalities. It is not unlikely that a firm selects particularly aggressive individuals if it intends to act aggressively in the market. Moreover, firms as corporate entities may themselves, in different degrees, care about relative, not only about absolute payoff. One reason is the embeddedness of some firms into financial markets, possibly also into a market for corporate control. In these markets, comparative performance may be a very relevant signal, whereas in other markets that might be less so.

That said, an experiment will not be able to settle the policy debate over mavericks. Experiments are only tools for identifying potential effects. But we add an

important argument to this policy debate. Maverick choices may be expected, they may be sustainable, and they may affect market outcomes, even in the absence of a pecuniary incentive to act aggressively. Anti-trust authorities have no reason to stop searching for, or protecting, maverick behavior, even if it does not seem to be grounded in sound profit incentives of the firm in question.

A second finding is of even greater importance for anti-trust policy: maverick behavior is not to be expected irrespective of context. When they face tough competition, even individuals (firms) otherwise inclined to compete aggressively are likely to hold back. We have of course only shown this for maverick behavior resulting from rivalistic preferences. But one should a fortiori expect a disciplining effect of a competitive environment on mavericks that have an incentive to outperform others (for instance since their income is tied to market share): by definition, maverick behavior reduces absolute profit. For anti-trust, this insight matters in merger control. Not so rarely, mergers between conglomerate firms reduce competition in one, but increase competition in another market. In principle, it makes sense to balance out these effects. But if the merger enables entry into a new market and competition in this market is intense, the merger is unlikely to increase consumer welfare, even if the entrant has an incentive to bid aggressively.

Appendix I: Model

In the general case of a Cournot market with linear demand, intercept m, and n suppliers, all with marginal cost of zero, the Cournot-Nash quantity is given by:

$$q_i = \frac{m}{n+1}$$

We now assume that the utility of the rivalistic supplier e (given that the other two suppliers make identical profits π_i, which will be the case in equilibrium) is given by

$$u_e = \pi_e + (n-1)\gamma(\pi_e - \pi_i)$$

$$= (1+2\gamma)(m-(n-1)q_i - q_e)q_e - 2\gamma(m-(n-1)q_i - q_e)q_i$$

Profit for one of the incumbents is now given by

$$\pi_i = \left(m - q_i - (n-2)q_j - q_e\right)q_i$$

Taking first order conditions, and solving the resulting system of equations, we get

$$q_i = q_j = \frac{m(2\gamma+1)}{2\gamma n + n + 1 + 4\gamma}, q_e = \frac{m(4\gamma+1)}{2\gamma n + n + 1 + 4\gamma}$$

E.g., with the parameters of the experiment, and letting the entrant be mildly rivalistic, i.e. with $\gamma = \frac{1}{2}$, we get $q_i = q_j = 22.22, q_e = 33.33$. The rivalistic player is better off

the larger γ, that is the more she is rivalistic. If all sellers hold standard preferences, in equilibrium they sell $Q_N = nq_i = \frac{nm}{n+1}$ units. If one seller is rivalistic, total quantity is given by

$$Q_R = (n-1)\frac{m(2\gamma+1)}{2\gamma n + n + 1 + 4\gamma} + \frac{m(4\gamma+1)}{2\gamma n + n + 1 + 4\gamma}$$

which is larger than Q_N for any $\gamma > 0$; with $\gamma = 0, Q_R = Q_N$. Hence consumer welfare increases if there is a rivalistic player.

Qualitatively similar results are obtained if we also allow incumbents to be rivalistic as shown below, if we keep the assumption that the entrant is more rivalistic $(\gamma_e \geq \gamma_i)$.[29] Specifically, let us assume that $\alpha = \gamma_i < \gamma_e = \gamma$. Taking first order conditions, and solving the resulting system of equations, we get

$$q_i = q_j = \frac{m(4\alpha\gamma + 2\alpha + 2\gamma + 1)}{4\alpha\gamma n + 2\alpha n + 2\gamma n + n + 1 + 4\gamma},$$

$$q_k = \frac{m(4\alpha\gamma + \alpha + 4\gamma + 1)}{4\alpha\gamma n + 2\alpha n + 2\gamma n + n + 1 + 4\gamma}$$

Similar to our previous results, each incumbent sells less than the entrant, and consumer welfare increases both in α and γ.

Our hypotheses are based on the assumption that preferences are common knowledge, and that all suppliers maximize utility. This is not what we find in the experiment. Visibly many duopolies are out of equilibrium, and entrants react to this. We therefore also report best responses of entrants, assuming that incumbents will only adjust quantities to the entry of one more supplier (i.e. will choose $q_3 = .75^* q_2$, where numbers 2 and 3 stand for the number of suppliers). If the entrant maximizes profit (is selfish), she will then choose the following best response

$$q_{e\_br} = \frac{1}{2}(m - (n-1).75^* q_2)$$

or, with the parameters of the experiment, $50 - .75^* q_2$. Note that, if the duopoly was in equilibrium, $75^* q_2 = 25$, so that the best response is the equilibrium. Hence the

29 In fact, the result can be generalized by noting that our model is related to the model by Fehr and Schmidt (1999). The difference is that the Fehr-Schmidt model allows players to also suffer from advantageous inequality. However, as long as the entrant is assumed to be more aggressive than the incumbents, the incumbents will in equilibrium always fall behind the entrant and so never experience advantageous inequality. Since the utility from the difference between one's own payoff and the payoff of a peer is not constrained to positive differences, our utility also captures disutility from falling behind one's peers. So, technically, this leads to a market of n players who all hold preferences as we assume above.

model predicts that entrants choose a larger quantity only if the duopoly was collusive. This fits the data from the *Zero* treatment very well, Figure 16.4.

If entrants are rivalistic, the best response to the expectation that incumbents will only adjust to the fact that one more supplier is in the market is given by maximizing the utility, assuming $q_3 = .75^*q_2$. In generic notation the best response is given by

$$q_{e\_br} = \frac{m + (n-1).75^*q_2 + (n-1)\gamma(m - (n-2).75^*q_2)}{2 + (2n-2)\gamma}$$

With the parameters of the experiment, this simplifies to

$$q_{e\_br} = \frac{50 - .75^*q_2 + (100 - .75^*q_2)\gamma}{1 + 2\gamma}$$

Note that this quantity is *below* the Nash quantity for large q_2 and/or for small γ. This fits the results from Figure 16.4 very well.

Appendix II: Regression equations

Table 3 model 1	$quantity_{it\|t>10} = \beta_0 + \beta_1 * treat + \beta_2 * mean\left(quantity_{g\|t<11}\right) + \beta_3 * treat * mean\left(quantity_{g\|t<11}\right) + \varepsilon_i + \varepsilon_{it}$
Table 3 model 2	$quantity_{it\|t>10} = \beta_0 + \beta_1 * treat + \beta_2 * mean\left(quantity_{g\|t<11}\right) + \beta_3 * treat * mean\left(quantity_{g\|t<11}\right) + \beta_4 * entrant + \beta_5 * treat * entrant + \beta_6 * entrant * mean\left(quantity_{g\|t<11}\right) + \beta_7 * entrant * treat * mean\left(quantity_{g\|t<11}\right) + \varepsilon_i + \varepsilon_{it}$
Table 3 model 3	$quantity_{it\|t>10} = \beta_0 + \beta_1 * treat + \beta_2 * mean\left(quantity_{g\|t<11}\right) + \beta_3 * treat * mean\left(quantity_{g\|t<11}\right) + \beta_4 * entrant + \beta_5 * treat * entrant + \beta_6 * entrant * mean\left(quantity_{g\|t<11}\right) + \beta_7 * entrant * treat * mean\left(quantity_{g\|t<11}\right) + \beta_8 * entrant'sSVO + \beta_9 * entrant'sSVO * mean\left(quantity_{g\|t<11}\right) + \varepsilon_i + \varepsilon_{it}$
Table 4 model 1	$\left(quantity_{it} - quantity_{i,t-1}\right)_{\|t>10,\, incumbent} = \beta_0 + \beta_1 * quantity_{t-1,\, entrant} + \beta_2 * quantity_{t-1,\, entrant} * mean\left(quantity_{g\|t<11}\right) + \varepsilon_i + \varepsilon_{it}$
Table 4 model 2	$\left(quantity_{it} - quantity_{i,t-1}\right)_{\|t>10,\, incumbent} = \beta_0 + \beta_1 * quantity_{t-1,\, entrant} + \beta_2 * quantity_{t-1,\, entrant} * mean\left(quantity_{g\|t<11}\right) + \beta_3 * quantity_{t-1,\, entrant} * mean\left(sd\left(mean\left(quantity_{g\|t<11}\right)\right)\right) + \beta_4 * quantity_{t-1,\, entrant} * mean\left(quantity_{g\|t<11}\right) * mean\left(sd\left(mean\left(quantity_{g\|t<11}\right)\right)\right) + \varepsilon_i + \varepsilon_{it}$

Appendix III: Instructions

a) Instructions: First session

(1) General instructions

Thank you for taking part in our experiment. From your invitation you already know that the experiment is in two parts. These instructions explain the first part of the experiment, taking place today. We will pay you your earnings from today's part of the experiment at the end of today's session. However, it is very important for our experiment that you also participate in the second session.

You can earn money in this experiment. How much you earn depends on your decisions and the decisions of other participants. Your earnings will be paid to you in cash at the end of the experiment.

Please switch off your mobile phone now, and please do not communicate any longer with the other participants as of this moment. Should you have a question about the experiment, please raise your hand. We will come to you and answer your query.

Today's part of the experiment consists of different sections. In these instructions, we explain the first section. For the following sections, you will find your instructions on the screen in front of you.

In order for us to keep track of your performance in the second part of the experiment, we would ask you please to generate an identification code at the end of the experiment, and to enter this code on your computer screen. We will use this identification code to connect your data from the first and second parts of the experiment. At no time do we know your name or address. Only the laboratory administration has that information. However, the laboratory administration does not know your decisions. This way we can ensure that **anonymity is guaranteed at all times**. Please write down this number and bring it with you when you are invited to the second experiment. At the beginning of the the second experiment, we will ask you to enter this number on your computer screen. **If you enter the wrong number, you cannot take part in the second experiment. Therefore, please check whether you have made a note of the correct number.**

(2) First section

We are now going to ask you to make several decisions. For this to happen, you will be randomly matched with another participant. You can allocate Taler to this participant and to yourself in the course of several distribution decisions. In order to do this, you will have to choose repeatedly between two distributions, X and Y (e.g., distribution X: 10 Taler for yourself and 12 Taler for the other player; and distribution Y: 8 Taler for yourself and 20 Taler for the other player). The Taler you allocate to yourself are paid out to you at the end of the experiment, at a rate of **100 Taler = 1 €**. At the same time, you are also randomly matched with yet another experiment participant who, in turn, can allocate Taler by way of distribution decisions.

This participant is **not the same** as the one to whom you can allocate Taler. The Taler allocated to you are also transferred to your account and paid out to you at the end of the experiment, at a rate of 100 Taler = 1 €.

The individual decision tasks will look like this:

[translation of screenshot
Period 1 of 1
Task
Please choose your preferred distribution of Taler.
Possibility A
Possibility B
Your Taler
The Taler of the participant matched with you]

b) Instructions: Second session

(1) General instructions

Welcome to the experiment! This is the second part of the experiment. The first part took place a few days ago. We would like to thank you for showing up once again. Please enter your identification number on your screen now. Let us remind you that we will not connect this number with your name and your address. You will therefore remain anonymous for both today's experiment and the earlier one. Your number will be used exclusively to relate your decisions from both experiments to you.

You can earn money in this experiment. How much you earn depends on your decisions and the decisions of other participants.

Please switch off your mobile phone now, and please do not communicate any longer with the other participants as of this moment. Should you have a question about the experiment, please raise your hand. We will come to you and answer your query.

This experiment is in three parts. You will find the instructions for the first part below. The instructions for the following parts will be handed out to you after the respective previous parts have been completed. As we will explain to you later on, participants can take on different roles in the course of the experiment.

Each of these parts consists of several rounds. All rounds of all parts are payoff-relevant. In this experiment, we use the Experimental Currency Unit ECU. All sums in ECU are always rounded off to whole numbers. At the end of the experiment, the sum of all ECU contributions is converted into Euro at a rate of **2000 ECU = 1 €.** The converted sum will be paid to you in cash at the end of the experiment.

You will remain in a group of three participants for the duration of the entire experiment. The constellation of the group does not change.

All decisions in this experiment, as well as the payoffs at the end, remain anonymous. Please do not discuss these with any of the other participants, even when the experiment has ended.

(2) Instructions: First part

CAUTION: One-third of the participants pauses in this part of the experiment and will not continue until the second part. However, these participants are also

informed about what is happening. We will inform you at the beginning of the experiment about the role you have in the first part.

This part of the experiment consists of 10 rounds. In each round, two participants are actors in a market. Both participants produce an identical product at no production costs. At the beginning of each round, each producer chooses the amount he or she wishes to produce. The market price (P), at which each unit is sold on the market, depends on the total amount (Q) produced by both participants. The market price is calculated as follows:

$$P = \begin{cases} 100 - Q & false\ Q < 100 \\ 0 & else \end{cases}$$

This means, first of all, that both producers receive the same market price for their amounts. Secondly, the higher the total amount Q is that both producers sell, the lower is the market price. As of a total amount of 100, the market price equals zero.

For each of the two producers, the payoff for the round is his or her chosen production amount, multiplied by the market price. The total payoff for this part of the experiment is the sum of all individual payoffs per round.

After each round, you will receive feedback on the amount the producers have chosen in total, on the market price, and on your earnings.

(3) Instructions: Second part

This part of the experiment consists of a 10-round market, just like the first part. The only difference now is that there is a further producer, in addition to the two "*older*" *producers*. The "*new producer*" has paused in the first part of the experiment, but received the same instructions as the two other producers, for the purpose of information. In addition, this new producer has also been informed about the market prices and amounts of the past ten rounds, concerning the group this new producer has joined.

Apart from the fact that there are now three producers, nothing else changes. As before, the market price is calculated for all three producers – the two old and the new – using the same formula:

$$P = \begin{cases} 100 - Q & if\ Q < 100 \\ 0 & else \end{cases}$$

This means all three producers receive the same market price P for their amounts, and that the market price that can be attained falls proportionally to the total amount Q rising.

(4) Instructions: Third part

This part of the experiment consists of a further continuation of the market by an additional ten rounds. However, both the two old producers who were active in the first part and the new producer who joined the market in the second part have the

opportunity to negotiate a possible departure of the new producer from the market. Negotiations are conducted according to the following rules.

Independently of the second producer, each of the two old producers names a *maximum* price figure, in ECU, which he or she would pay the new producer if this producer were prepared, in return, to quit the game for the additional ten rounds. However, the highest possible price that the two old producers can name is the figure you have earned in the first two parts of the experiment.

At the same time, the new producer names a figure *B* (in ECU), beginning with which he or she is willing to forfeit participation in the additional ten market rounds.

Then, one of the two offers made by the old producers is chosen randomly, with each offer having a 50-percent chance of being chosen. There are two possibilities:

- If the maximum offer *A* of the old producer who has been chosen is at least as high as the new producer's demand *B*, then the old producer who has been chosen pays the new producer demand *B*. (Offer *A* hence describes the chosen old producer's *maximum* willingness to pay; usually, less is paid.) Then, the additional ten market rounds take place *without* the new producer – as in the first part of the experiment.
- If the maximum offer *A* of the old producer who has been chosen is smaller than the new producer's demand *B*, then the additional ten market rounds take place *with* the new producer – as in the second part of the experiment. In this case, there is no exchange of any payment between the chosen old and the new producer.

References

Auriol, Emmanuelle, and Régis Renault. 2008. "Status and Incentives." *Rand Journal of Economics* 39 (1):305–326.

Baker, Jonathan B. 2002. "Mavericks, Mergers, and Exclusion: Proving Coordinated Competitive Effects Under the Antitrust Laws." *New York University Law Review* 77:135–203.

Ball, Sheryl, and Catherine C. Eckel. 1998. "The Economic Value of Status." *Journal of Socio-Economics* 27 (4):495–514.

Ball, Sheryl, Catherine Eckel, Philip J. Grossman, and William Zame. 2001. "Status in Markets." *The Quarterly Journal of Economics* 116 (1):161–188.

Bazerman, Max H., George F. Loewenstein, and Sally Blount White. 1992. "Reversals of Preference in Allocation Decisions. Judging an Alternative versus Choosing among Alternatives." *Administrative Science Quarterly* 37:220–240.

Boes, Stefan, Kevin Staub, and Rainer Winkelmann. 2010. "Relative Status and Satisfaction." *Economics Letters* 109 (3):168–170.

Bolton, Gary E. 1991. "A Comparative Model of Bargaining. Theory and Evidence." *American Economic Review* 81:1096–1136.

Bolton, Gary E., and Axel Ockenfels. 2000. "ERC: A Theory of Equity, Reciprocity and Competition." *American Economic Review* 90:166–193.

Breunig, Robert, and Flavio Menezes. 2008. "Empirical Approaches for Identifying Maverick Firms. An Application to Mortgage Providers in Australia." *Journal of Competition Law and Economics* 4 (3):811–836.

Brosig, Jeannette, Joachim Weimann, and Axel Ockenfels. 2003. "The Effect of Communication Media on Cooperation." *German Economic Review* 4:217–241.

Busse, Meghan. 2002. "Firm Financial Condition and Airline Price Wars." *Rand Journal of Economics* 33:298–318.

Camerer, Colin F. 2003. *Behavioral Game Theory. Experiments in Strategic Interaction*. 1. print. ed, *The roundtable series in behavioral economics*. New York: Sage.

Charness, Gary, David Masclet, and Marie Claire Villeval. 2013. "The Dark Side of Competition for Status." *Management Science* 60 (1):38–55.

Chaudhuri, Ananish. 2011. "Sustaining Cooperation in Laboratory Public Goods Experiments. A Selective Survey of the Literature." *Experimental Economics* 14:47–83.

Clark, Andrew E., Paul Frijters, and Michael A. Shields. 2008. "Relative Income, Happiness, and Utility. An Explanation for the Easterlin Paradox and Other Puzzles." *Journal of Economic Literature* 46:95–144.

Clemens, Christiane. 2006. "Status Concerns and Occupational Choice under Uncertainty." *Advances in Theoretical Economics* 6 (1):1–25.

Cooper, David J, and John Kagel. 2016. "Other-regarding Preferences." In *Handbook of Experimental Economics*, edited by John H. Kagel and Alvin E. Roth, 217–289. Princeton University Press, Princeton, NJ.

Cramton, Peter, and Axel Ockenfels. 2017. "The German 4G German Auction: Design and Behavior." *The Economic Journal* 127: F305–F324.

Crandall, Robert W., and Clifford Winston. 2003. "Does Antitrust Policy Improve Consumer Welfare? Assessing the Evidence." *Journal of Economic Perspectives* 17:3–26.

De Dreu, Carsten K. W., and Terry L. Boles. 1998. "Share and Share Alike or Winner Take All? The Influence of Social Value Orientation upon Choice and Recall of Negotiation Heuristics." *Organizational Behavior and Human Decision Processes* 76 (3):253–276.

Eckert, Andrew, and Douglas S. West. 2004a. "Retail Gasoline Price Cycles across Spatially Dispersed Gasoline Stations." *Journal of Law and Economics* 47 (1):245–273.

Eckert, Andrew, and Douglas S. West. 2004b. "A Tale of Two Cities. Price Uniformity and Price Volatility in Gasoline Retailing." *Annals of Regional Science* 38 (1):25–46.

Engel, Christoph. 2007. "How Much Collusion? A Meta-Analysis on Oligopoly Experiments." *Journal of Competition Law and Economics* 3:491–549.

Engel, Christoph. 2010. "The Behaviour of Corporate Actors. A Survey of the Empirical Literature." *Journal of Institutional Economics* 6:445–475.

Engel, Christoph. 2015. "Tacit Collusion. The Neglected Experimental Evidence." *Journal of Empirical Legal Studies* 12:537–577.

Engel, Christoph. 2016. "A Random Shock is not Random Assignment." *Economics Letters* 145: 45–47.

Farrell, Joseph, and Paul Klemperer. 2007. "Coordination and Lock-in. Competition with Switching Costs and Network Effects." In *Handbook of Industrial Organization*, edited by Mark Armstrong and Robert H. Porter, 1967–2072. Amsterdam: Elsevier.

Fehr, Ernst, and Klaus M. Schmidt. 1999. "A Theory of Fairness, Competition, and Cooperation." *Quarterly Journal of Economics* 114:817–868.

Ferrer-i-Carbonell, Ada. 2005. "Income and Well-being. An Empirical Analysis of the Comparison Income Effect." *Journal of Public Economics* 89 (5):997–1019.

Fischbacher, Urs. 2007. "z-Tree. Zurich Toolbox for Ready-made Economic Experiments." *Experimental Economics* 10:171–178.

Fouraker, Lawrence E., and Sidney Siegel. 1963. *Bargaining Behavior*. New York: McGraw-Hill.
Frank, Robert H. 1984. "Are Workers Paid Their Marginal Products?" *American Economic Review* 74:549–571.
Frank, Robert H. 1985. *Choosing the Right Pond. Human Behavior and the Quest for Status*: Oxford University Press. New York and Oxford.
Goppelsroeder, Marie. 2009. Three is still a Party. An Experiment on Collusion and Entry.
Greiner, Ben. 2004. "An Online Recruiting System for Economic Experiments." In *Forschung und wissenschaftliches Rechnen 2003*, edited by Kurt Kremer and Volker Macho, 79–93. Göttingen: GWDG.
Grund, Christian, and Johannes Martin. 2017. "Monetary Reference Points of Managers–Empirical Evidence of Status Quo Preferences and Social Comparisons." *Scottish Journal of Political Economy* 64 (1):70–87.
Huberman, Bernardo A., Christoph H. Loch, and Ayse Önçüler. 2004. "Status as a Valued Resource." *Social Psychology Quarterly* 67 (1):103–114.
İriş, Doruk, and Luís Santos-Pinto. 2014. "Experimental Cournot Oligopoly and Inequity Aversion." *Theory and Decision* 76 (1):31–45.
Klemperer, Paul. 2004. *Auctions. Theory and Practice*, *The Toulouse Lectures in Economics*. Princeton, N.J.: Princeton University Press.
Kolasky, William J. 2002. Coordinated Effects in Merger Review. From Dead Frenchmen to Beautiful Minds and Mavericks.
Ku, Gillian, Deepak Malhotra, and J. Keith Murnighan. 2005. "Towards a Competitive Arousal Model of Decision-making. A Study of Auction Fever in Live and Internet Auctions." *Organizational Behavior and Human Decision Processes* 96 (2):89–103.
Kwoka, John E. 1989. "The Private Profitability of Horizontal Mergers with non-Cournot and Maverick Behavior." *International Journal of Industrial Organization* 7 (3):403–411.
Li, Jin, and Charles R. Plott. 2009. "Tacit Collusion in Auctions and Conditions for its Facilitation and Prevention. Equilibrium Selection in Laboratory Experimental Markets." *Economic Inquiry* 47 (3):425–448.
Liebrand, Wim B., and Charles G. McClintock. 1988. "The Ring Measure of Social Values. A Computerized Procedure for Assessing Individual Differences in Information Processing and Social Value Orientation." *European Journal of Personality* 2:217–230.
Loch, Christoph H., and Yaozhong Wu. 2008. "Social Preferences and Supply Chain Performance. An Experimental Study." *Management Science* 54 (11):1835–1849.
Luttmer, Erzo F. P. 2005. "Neighbors as Negatives. Relative Earnings and Well-being." *Quarterly Journal of Economics* 120 (3):963–1002.
Malhotra, Deepak. 2010. "The Desire to Win. The Effects of Competitive Arousal on Motivation and Behavior." *Organizational Behavior and Human Decision Processes* 111 (2):139–146.
Malhotra, Deepak, Gillian Ku, and J. Keith Murnighan. 2008. "When Winning is Everything." *Harvard Business Review* 86 (5):78.
Messick, David M., and Warren B. Thorngate. 1967. "Relative Gain Maximization in Experimental Games." *Journal of Experimental Social Psychology* 3 (1):85–101.
Ockenfels, Axel, Dirk Sliwka, and Peter Werner. 2015. "Bonus Payments and Reference Point Violations." *Management Science* 61 (7):1496–1513.
Scherer, Frederic M., and David Ross. 1990. *Industrial Market Structure and Economic Performance*. 3. ed. Boston: Houghton Mifflin.
Sobel, Joel. 2009. "Generous Actors, Selfish Actions. Markets with Other-regarding Preferences." *International Review of Economics* 56 (1):3–16.

Solnick, Sara J., and David Hemenway. 1998. "Is More Always Better?: A Survey on Positional Concerns." *Journal of Economic Behavior and Organization* 37 (3):373–383.

Van Damme, Eric. 2003. "The Dutch UMTS Auction." In *Spectrum Auctions and Competition in Telecommunications*, edited by Gerhard Illing and Ulrich Klüh, 263–294. München: CES Ifo.

Van Lange, Paul A. M., Ellen De Bruin, Wilma Otten, and Jeffrey A. Joireman. 1997. "Development of Prosocial, Individualistic, and Competitive Orientations. Theory and Preliminary Evidence." *Journal of Personality and Social Psychology* 73 (4):733–746.

Rense Corten, Vincent Buskens and Stephanie Rosenkranz

17 Cooperation, Reputation Effects, and Network Dynamics: Experimental Evidence

Abstract: While social network structures are thought to promote cooperation through reputation effects, as suggested by Raub and Weesie (1990), the option of partner choice may undermine these reputation effects in networks. This article approaches this dilemma by comparing the effects of partner choice and reputation diffusion in isolation as well as in combination in a controlled experimental setting. While we do not find that cooperation rates in the absence of partner choice are higher in the presence of reputation effects, we find that emerging cooperation levels near the end of the game are higher when initial cooperation levels are higher. This is more in line with predictions of models of cooperation that rely on learning heuristics rather than forward-looking rationality (i.e., Corten and Cook, 2009). Moreover, we find that the option of partner choice lowers cooperation rates in the absence of reputation effects. However, we do not find a similar effect in the presence of reputation effects. We position these findings in the larger literature on the conditions for cooperation in dynamic societies.

17.1 Introduction

Cooperation is a cornerstone of human societies (e.g., Bowles and Gintis 2013; Pennisi 2005; Ostrom 1990). In many instances of social interaction, people join forces to achieve something they could not have achieved alone. Achieving cooperation, however, is often problematic: actors may face incentives to free-ride on the efforts of others, with the result that cooperation never materializes and the payoff to all actors involved is lower than it would have been, had they cooperated. Consider, for example, two researchers who can collaborate on a project, but are also tempted to let the other do most of the work and focus on their individual projects. This situation is formally captured for two actors in the famous Prisoner's Dilemma (PD).[1] The question as to under which conditions cooperation between rational, selfish actors becomes more likely is one of the major problems of the social sciences, and is also known in sociology as the problem of social order (Parsons, 1937). A key finding in this line of research is that cooperation is possible if interactions are repeated (Axelrod, 1981; Axelrod and Hamilton 1981; Taylor, 1977). Raub

1 Many more examples of social dilemma problems and especially in the academic world have vividly been illustrated by Raub in his farewell lecture (Raub 2017: 49–54).

Rense Corten, Vincent Buskens, Stephanie Rosenkranz, Utrecht University, Utrecht

https://doi.org/10.1515/9783110647495-017

and Voss were among the first to formalize this dilemma game-theoretically within sociology (e.g., Voss 1982; Raub and Voss 1986a) following Coleman's (1964: 167) advice to start from "an image of man as wholly free: unsocialized, entirely self-interested, not constrained by norms of a system, but only rationally calculating to further his own self-interest." However, the assumptions under which this result was initially obtained were rather restricted. Consequently, scholars have searched for additional mechanisms that facilitate the emergence of cooperation.

A key assumption in the 'baseline scenario' of repeated interaction (in addition to the rationality assumptions) is that interactions occur in social isolation. Actors interact only with one partner at a time and have no information on interactions in which they are not involved. In reality, however, cooperative relations are often embedded in social networks, through which information on what happens in one interaction becomes known to third parties (Granovetter 1985). An intuitive and broadly shared view among social scientists is that in such 'embedded scenarios' the emergence of cooperation is more likely (Homans 1958; Coleman 1990; Raub and Voss 1986b; Voss, 2001), a view supported by much qualitative (Macaulay 1963; Greif 1989, 1994; Ellickson 1991; Uzzi 1996, 1997) and some quantitative evidence (e.g., Burt and Knez 1995; Buskens 2002; Raub and Buskens 2013). In our example, cooperation in common research projects would be more likely in departments with dense networks, in which information about defections is easily shared among colleagues. This information can impact cooperation in social dilemmas through reputation effects. Actors embedded in networks may be more reluctant to defect because word regarding their behavior will spread and lead to sanctions by third parties. In probably the first paper modelling games on networks, Raub and Weesie (1990) show that such reputation effects indeed render conditional cooperation by selfish and rational actors more likely. Moreover, actors may learn from previous experiences that cooperation with certain partners is more profitable (Buskens and Raub 2002).

A second restrictive assumption in the 'baseline scenario' is that actors are forced to stay in interactions with a given partner, that is, they do not have opportunities to voluntarily enter and leave interactions. Theoretical studies show that in scenarios in which actors do have such opportunities for partner choice, cooperation can be sustained (Schuessler 1989), or may even increase as compared to the situation in which actors are forced to play with a certain partner, because cooperative actors can avoid defectors (Stanley et al. 1995; Hauk 2001). Experimental studies (Orbell and Dawes 1993; Boone and Macy 1999; Hauk and Nagel 2001) corroborate this claim. In our example, this theory would predict that cooperation in common research projects is more likely if researchers can freely choose their coauthors and can ostracize free-riding colleagues.

A third restrictive assumption in the 'baseline scenario' is that actors cannot impose formal arrangements or add other institutional arrangement to enforce cooperation (e.g., include options of peer punishment as in Fehr and Gächter 2002). We do not consider extensions along this third assumption in this paper.

The above results on the impact of social embeddedness and partner choice are important, but to some extent limited. Many social situations combine social networks and partner choice. When interactions are embedded in a social network, actors often have possibilities to (at least to some degree) choose their relationships in the network and, conversely, situations in which actors can choose partner, are often embedded in social networks. Our example above, in which both mechanisms are plausible, clearly illustrates this point. We therefore propose a scenario in which interactions are embedded in a social network in the sense that information about interactions is diffused through the network of interactions, but actors also have opportunities to change the network by choosing interaction partners. This scenario can be characterized as a dynamic social network. Although the effects of social networks and partner choice on cooperation are relatively well understood, we know little about the combination in dynamic social networks. The proposed research aims to fill this gap by systematically comparing the above scenarios in laboratory experiments.

Predicting the combined effects of partner choice and social networks is not trivial (see Corten and Cook 2009; Raub, Buskens, and Frey 2013). Consider the following intuitive but contradictory hypotheses on these effects. On the one hand, we may expect that the positive effects of embeddedness and partner choice on cooperation simply add up to produce even higher levels of cooperation, or even reinforce each other, as the diffusion of reputations leads to more effective partner choice.

On the other hand, one may argue that the effects of network reputation and partner choice undermine each other, leading to lower levels of cooperation. First, it is possible that the reputation effects that promote cooperation in social networks are corroded by the possibility of partner choice: if actors can modify the network, this might (unintentionally) break information channels necessary for reputation building and learning. Second, reputation effects in networks might interfere with partner choice in ways that are detrimental to cooperation. For instance, an actor who somehow earned a bad reputation as defector may have difficulties finding new interaction partners, even if she has changed her behavior to cooperation. Conversely, actors may be reluctant to terminate interactions with partners who earned a favorable reputation in the past, even if they start to defect.

This paper seeks to explore these competing mechanisms by deriving specific hypotheses and testing these hypotheses empirically through laboratory experiments. The experiments systematically compare effects on cooperation of four different treatments that correspond with the four scenarios sketched above:

1. No reputation effects and no partner choice (the ‘Baseline Treatment’);
2. Reputation effects without partner choice (the ‘Reputation Treatment’);
3. Partner choice without reputation effects (the ‘Partner Choice Treatment’);
4. Reputation effects with partner choice (the ‘Reputation-Partner Choice Treatment’).

Previous research suggests that cooperation levels in both the Reputation Treatment as well as in the Partner Choice Treatment are higher than in the Baseline Treatment. The intuitive arguments sketched above, however, show that it is far from clear whether cooperation levels in the Reputation-Partner Choice Treatment are higher or lower than in the Reputation or Partner Choice Treatment. Because the mechanisms of reputation in networks and partner choice are likely to occur together in natural social situations, it is difficult to disentangle the two in observational research in field settings. Experiments, on the other hand, allow for studying the effects of reputation and partner choice both in isolation and in combination because the availability of the two mechanisms can be manipulated by the experimenter, while keeping other conditions constant (Falk and Heckman 2009).

Similar experiments that combine social dilemmas with dynamic networks have been conducted by Corbae and Duffy (2008) and Corten and Buskens (2009), who study Coordination Games rather than PDs. Close to our experiment is the experimental work by Ule (2005), who studies N-person PD's (i.e., subjects choose Cooperation or Defect, but play the same behavior with all their neighbors) instead of networked 2-person PD's (i.e., subjects chooses Cooperation or Defect with each neighbor separately), which changes the strategic decision problem considerably. Recently, theoretically as well as experimentally, the question of cooperation on networks has gained a lot of attention also from physicists leading to a hose of simulation models and some large-scale experiments. Casella et al. (2018; see also Sanchez 2018) provide a nice and concise review of this literature. The overview shows that theoretically, the relation between cooperation and network structure largely depends on details of the model. Experimental results suggest at best limited effects of network structure on cooperation in static networks (e.g., Cassar, 2007; Corten et al. 2016; Gracia-Lázaro et al. 2012; Grujić et al. 2010; Kirchkamp and Nagel 2007; Rand et al. 2014; Sanchez 2018), while dynamic networks seem to be more successful in sustaining cooperation (Rand et al. 2011). Note that also in these experiments, subjects play N-person PD's rather than several 2-person PDs. Recently, Harrell et al. (2018) confirmed the positive effect of reputation for the situation where participants interact in several 2-person PDs in a partially static network where some of the ties in the network can be changed.

In this paper we present a systematic test comparing cooperative behavior in the four treatments presented above in which subjects play 2-person PDs in (dynamic) networks. We start by formulating hypotheses on the expected differences between the treatments. Although we advocate rigorous theory development (see Raub and Buskens 2011; Raub 2017) and prefer formal derivation of hypotheses using a micro-macro perspective (Raub and Voss 1981, 2016; Coleman 1986, 1990; Raub et al. 2011), we defer to a more intuitive way of hypothesis formulation here based on more formal arguments given elsewhere. Thereafter, we present our experiment and the analysis of behavior. We end with conclusions and some reflections of the strengths and weaknesses of our study.

17.2 Theory

17.2.1 Reputation effects in exogenous networks

The most straightforward prediction is that we expect reputation to promote cooperation in the situation in which actors cannot choose their partners. This follows directly from, among others, the game-theoretic model by Raub and Weesie (1990). They show that in an infinitely repeated PD in which there is also information exchange between actors in different games, the more extensive this information exchange is, the less restrictive the conditions become under which an equilibrium is possible in which actors always cooperate. Under the interpretation that under information exchange the situation is more favorable for cooperation, the following hypothesis can be formulated:

Hypothesis 1: Cooperation is higher in the Reputation Treatment than in the Baseline Treatment.

Although this hypothesis seems intuitive and straightforward, we highlight some limitations of the model by Raub and Weesie (1990). The equilibrium conditions are based on the assumption that actors play trigger strategies, which presupposes that all actors are perfectly rational and coordinate on the cooperative equilibrium if it exists. This assumption is problematic because if actors would make mistakes and, e.g., defect while cooperation would still be the equilibrium behavior, these mistakes are likely to be more destructive if information exchange is more extensive. If others hear about the defections, they will react with defections themselves and the cooperative equilibrium collapses faster or more likely compared to the situation in which such reputation effect are not possible. Using a simulation model based on learning dynamics Cook and Corten (2009; see also Corten 2014: Ch. 3) show that in the Reputation Treatment more extreme levels (higher as well as lower) of cooperation are reached than in the Baseline Treatment. Thus, they show that the average level of cooperation does not necessarily increase. This might be a reason why many studies do not find a strong relation between possibilities for reputation building in networks and cooperation (see also Corten et al. 2016 in which we analyze part of the experiment for which we do the full analysis here, focusing on Hypothesis 1 in particular). Still, what Corten and Cook (2009) also show is that the variation in levels of cooperation achieved over groups increases with reputation and that this variation can be explained by initial tendencies to behave cooperatively or not. This is also consistent with other empirical observations related to network effects, for example, on trust (Burt and Knez 1995). These considerations motivate the following two hypotheses.

Hypothesis 2: Emerging cooperation levels are higher when initial cooperation levels are higher.

Hypothesis 3: The variance in cooperation is higher in the Reputation Treatment than in the Baseline Treatment.

17.2.2 Effects of partner choice without reputation

Under many conditions, it has been shown that adding an exit option to a PD, such that actors can avoid uncooperative subjects, increases the possibilities for cooperation (Schüssler 1989; Stanley et al. 1995; EDK-group 2000; Hauk 2001). However, the extent to which it helps depends on the attractiveness of the outside option and the way in which partnerships can be formed or partners can be avoided (e.g., Hauk and Nagel 2001; Hauk 2001). The extent to which partners can be punished by excluding them depends on how attractive the outside option is (that is, what the payoff is of not interacting with someone) and to what extent partners then want to use that option for punishment. In our set-up, the outside option is actually more attractive than mutual defection, while partners only play when they both are willing to play with each other. This largely resembles the mutual agreement scenario of Hauk and Nagel (2001) for which they find that defection rates are lower conditional upon both actors entering the PD if the outside option indeed exists. But also, the overall cooperation rate is lower in games with the outside option because there are quite some games in which actors end up in the outside option.

Strict game-theoretic predictions for the finitely repeated game with the attractive outside option would simply predict that actors play the outside option. In an infinitely repeated game, adding the attractive outside option should reduce the possibilities for cooperation because the sanction for uncooperative behavior becomes smaller. On the other hand, actors should only enter the game if they believe that the other actor will cooperate and hence if both consider the sanction possibilities within the PD strong enough for cooperation to be sustained. Based on these arguments, we reach the following hypothesis.

> *Hypothesis 4: Cooperation is higher in the Partner Choice Treatment than in the Baseline Treatment given that two actors do enter in an interaction.*

This is also consistent with a more recent experiment (Wilson and Wu 2017) that finds that cooperation increases with an outside option, but the likelihood to stay with a partner decreases with the attractiveness of the outside option. Given that cooperation within interactions is higher, while there will also be pairs of actors that do not interact, the difference between the Partner Choice Treatment and the Baseline Treatment in terms of total cooperation is uncertain.

17.2.3 Effects of partner choice in combination with reputation

There are few straightforward theoretical predictions that can be used to predict differences between the treatments with partner choice and reputation compared to the other three treatments. An exception is the study by Corten and Cook (2009) who

present a simulation in which the treaments of this experiment are combined (see also Fu, Hauert, and Nowak 2008 for a model version in which actors can switch partners but always keep the same number of interaction partners and can only choose the same behavior towards all their partners). We discuss some related research and experimental findings as a comparison. Riedl and Ule (2002) were among the first to study cooperation in dynamic networks. Their set-up is crucially different because they do not allow actors to differentiate behavior between partners. So actors can only choose to cooperate or defect with everyone or no one at the same time. This implies that cooperation is not determined at the dyadic level and observing cooperation tells something about the actor rather than about a specific relation. Therefore, their finding that cooperation is larger in dynamic than in fixed networks cannot be directly related to our experiment. Some later papers (Fehl et al. 2011; Rand et al. 2011) confirm the cooperation-promoting ability of dynamic networks in settings in which subjects cannot differentiate between their partners.

Strict game-theoretic predictions on models with network dynamics and reputation possibilities are hard to find and also difficult to generalize to different assumptions regarding incentives generated by the network (see Raub, Buskens and Frey 2013). Moreover, the equilibria derived in these models are typically based on trigger-strategy type of equilibrium analyses that are very sensitive to small deviations in behavior, especially if the information on these deviations can spread through the network. Although the conditions derived could still be interpreted in a more lenient way, namely, as how difficult it is to sustain cooperation under certain circumstances (see Raub and Buskens 2013: 125), we focus our predictions here on simulation analyses done to resemble the experiment we discuss.

Related simulations can also be found, e.g., by Melamed and Simpson (2016), which show that the emerging amount of cooperation might, in subtle ways, depend on the starting network structure, the value of ties compared to the payoffs in the games, and the speed of changing relations compared to the speed of adaptation of behavior. Relations should be able to be changed fast enough such that actors do not change to defective behavior because they are hooked up too long with other defectors. At the same time, relations should be stable enough to let cooperation be established within these relations. Melamed and Simpson (2016) test their model in an experiment in which participants play against pre-programmed opponents (while they are told that they play against real others). This, however implies that the results to a large extent are driven by the programmed strategies of opponents and can hardly be considered as a test of what would happen if real subjects interact. Later, in an M-Turk experiment the authors improved this set-up (Melamed et al. 2018), and added a set-up similar to ours in which subjects could differentiate their actions between others. They show that dynamic interaction dramatically increases cooperation even if when reputation effects are excluded (although it should be noted that Cuesta et al. 2018 claim that some reputation is still involved

in their no-reputation treatment). Still, their results show almost perfect cooperation levels in about all treatments with dynamic networks.

Considering the simulations of Corten and Cook (2009), which resemble our experimental set-up closely, the most salient observation is not that dynamics univocally promote cooperation, but that this finding strongly depends on the tendency to which actors start with the intention to cooperate or not. This is still consistent with other findings in which high cooperation levels are found, because in many subject pools cooperative tendencies are relatively high at the beginning of the experiment. However, it also suggests that if starting conditions vary considerably results might change dramatically. Therefore, we include hypotheses 2 and 3 regarding the importance of starting levels of cooperative behavior and the variance in cooperative behavior for dynamic networks.

Given earlier experimental evidence showing that starting levels of cooperative behavior can be expected to be relatively high, we expect that also with dynamic networks the effect of reputation will be positive overall. The simulations do not provide predictions on interaction effects from reputation and dynamic networks on levels of cooperation or on variations in cooperation.

17.2.4 Experimental setup

17.2.4.1 Treatments and the computer interface

The four treatments were implemented in a computer interface using z-Tree (Fischbacher 2007). Each experimental sessions consisted of 40 periods. We start with explaining the Baseline Treatment. In the Baseline Treatment, subjects interacted in groups of six. In each period, subjects were randomly matched with two other subjects in their group. They played a game with every other subject (see Figure 17.1). As Figure 17.1 shows, subjects received 30 points for every interaction that was not matched in any given period. This implies that in every period, subjects received 3 times 30 = 90 points, on top of the payoffs from matched interactions, regardless of their choices. The payoff for non-matched interactions was implemented for comparison with the two treatments in which subjects could choose partners, as described below. Because these payoffs do not in any way depend on the subjects' choices, they are not expected to influence the results in the treatments without partner choice.

Figure 17.2 shows the screen in which subjects made their decisions in the Baseline Treatment. The left-hand side of the screen represents the current choice situation. The yellow square represents the focal subject (Ego); the other subjects are represented by circles. The thin black lines between subjects indicate neighbors, i.e., all the relations that might be selected for an interaction in a given period in this group (in the Baseline Treatment, these were *all* dyads in the group). The black

		Player 2		
		BLUE	ORANGE	No Interaction
Player 1	BLUE	20,20	60,0	30,30
	ORANGE	0,60	40,40	30,30
	No Interaction	30,30	30,30	30,30

Figure 17.1: The experimental game.

circles represent the other subjects with whom Ego was matched for this period (subjects 4 and 5, in the example). This is furthermore indicated by the thick grey lines behind the thin black lines. By observing these thick grey lines, Ego can also learn which other pairs were matched in this period (in the example, these are 2 and 6, 2 and 3, 3 and 4, and 5 and 6).

The choices of the subjects are represented in the interface by arrows: if Ego chooses to play ORANGE (cooperation) against a partner, this is indicated by an orange arrow *from* Ego *to* this partner. Ego can indicate her choice by clicking with the cursor on the circles of the matched partners, which will change the color of the arrow. If Ego interacted before with any of her matched partners, the choices that were made in that previous interaction are already displayed on the screen and Ego can update her choice as desired. The upper right-hand corner of the screen shows the history of outcomes so far, which Ego can freely browse (using the 'next' and 'previous' buttons) for reference.

When Ego is satisfied with her choice, she clicks 'OK', which brings up the results screen shown in Figure 17.3. This screen shows the actions of Ego and her interaction partners and reports Ego's payoffs. In this example, Ego earned 40 points from the interaction with subject 5, 60 from the interaction with subject 4, and three times 30 for the other subjects with whom she did not interact, totaling 190 points. As in the choice screen, the upper right-hand corner of the screen provides the history of previous outcomes for reference.

The three other treatments differed from the Baseline Treatment by the implementation of one or both of two additional mechanisms: *reputation* and *partner choice*, which were implemented as follows.

In the treatments with reputation, subjects were informed not only about their own interactions but also about all interactions of their *neighbors*, which are all the other subjects with whom Ego could be matched (indicated by thin black lines on the screen). In our implementation, these were all subjects in the network in the

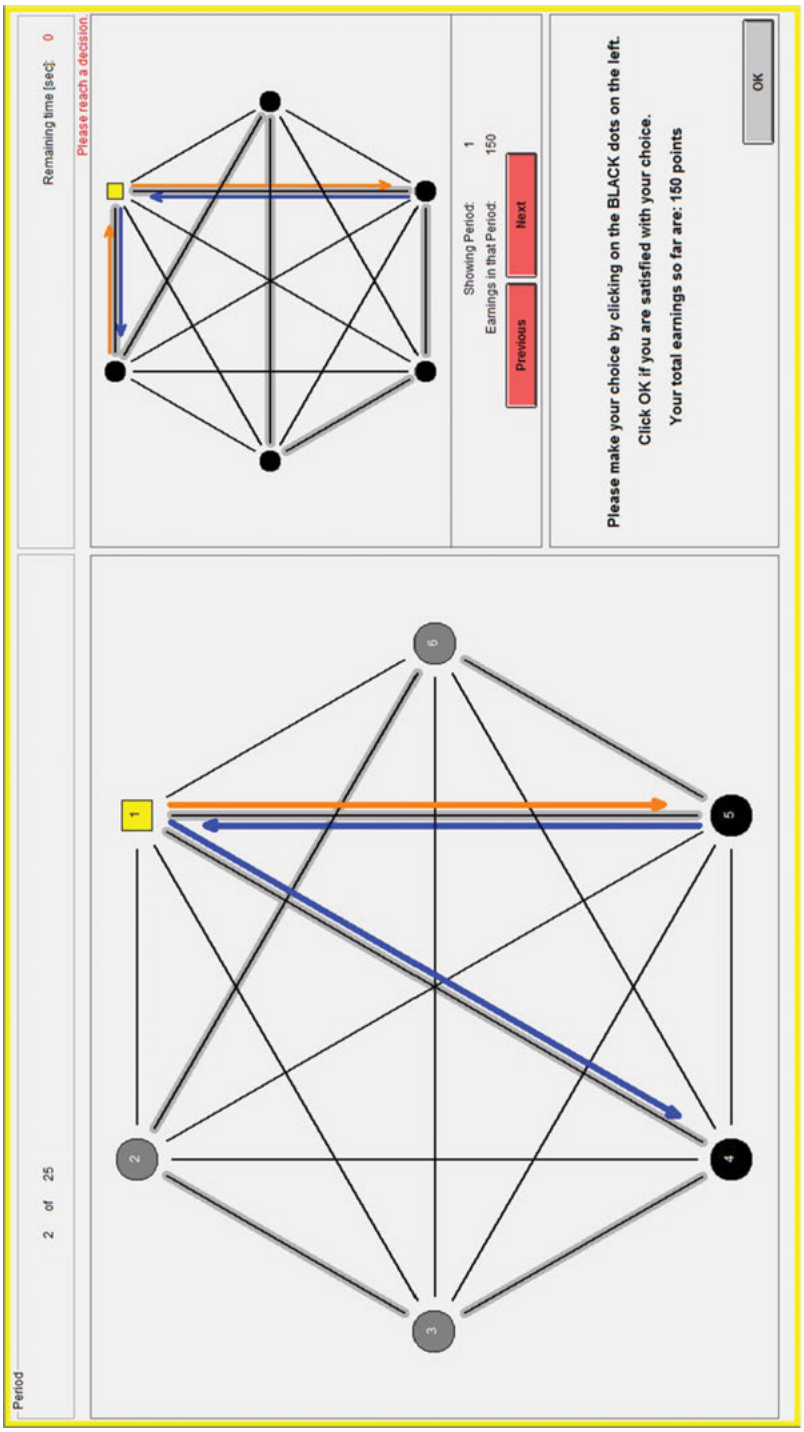


Figure 17.2: The computer interface: the choice screen of the Baseline Treatment.

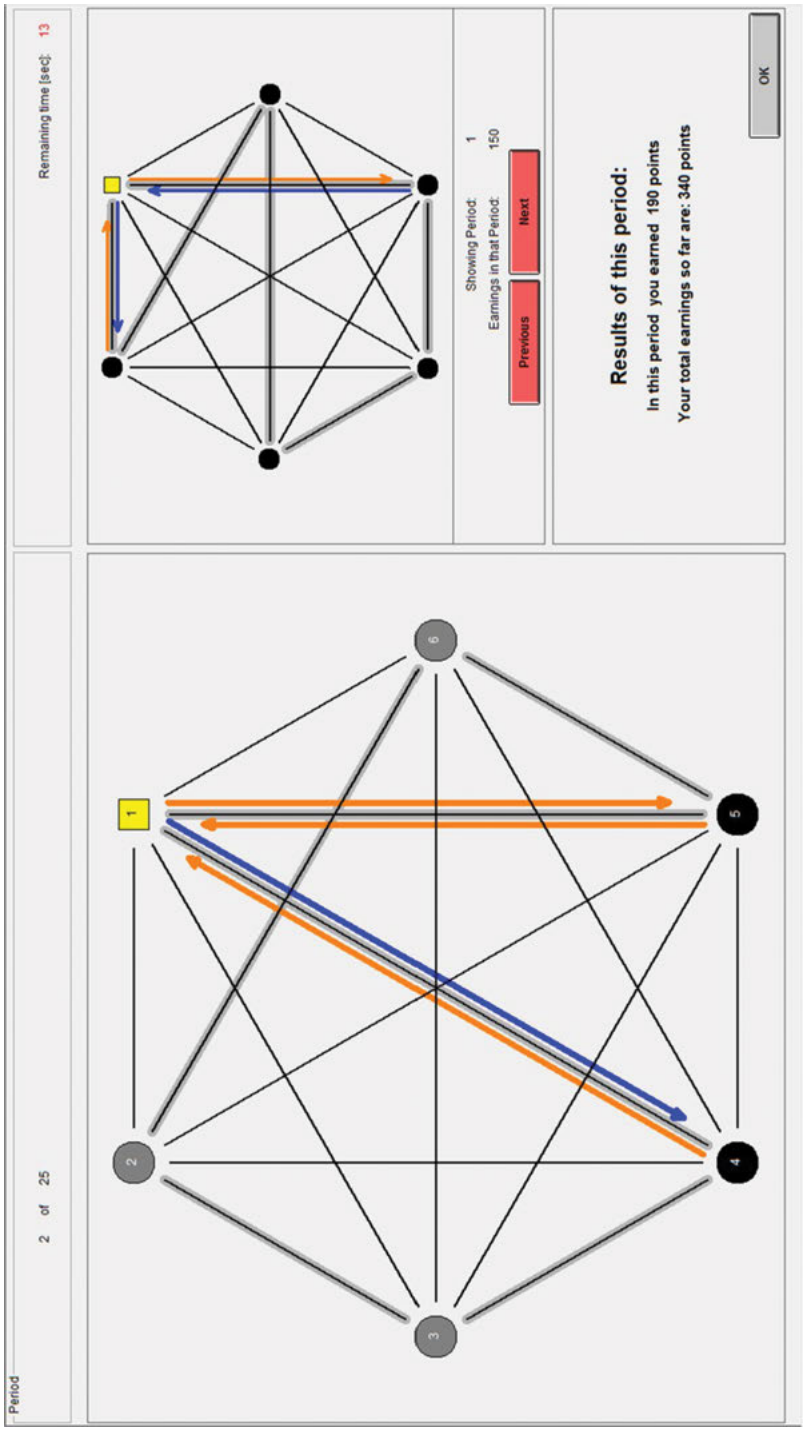


Figure 17.3: The computer interface: the results screen of the Baseline Treatment.

Reputation Treatment. In terms of information this implies that subjects were informed not only about the outcomes of their own interactions, but also about the outcomes of *all other* interactions. The interface of the Reputation Treatment differs from the Baseline Treatment only to the extent that outcomes of all other interactions are also displayed, as illustrated by Figure 17.4, which shows the results screen from the Reputation Treatment. Here, arrows are not only displayed for Ego's own interactions, but also for all other interactions that took place in that period.

In the treatments with partner choice, subjects had a third choice for every matched interaction besides ORANGE and BLUE, namely *not to interact* with this specific partner. This implies that if a pair of subjects in the given tie had been playing the PD in previous periods, each of them had the option to discontinue the interaction, while if they were not playing the game before, each could decide to start playing the PD. The subjects thus had the option to freely choose interaction partners, although this choice was restricted to other subjects you happened to be matched to for a potential interaction in a certain period. In this treatment, subjects could cycle though ORANGE, BLUE and NO INTERACTION by clicking on the circle representing the other subject. If one of the two subjects involved in the interaction chose NO INTERACTION, they both received the no interaction payoff of 30 (see Figure 17.1). In that case, they also were not considered neighbors, and had no thin black line between them in the interface. Still, everyone could be matched to everyone to play again and choose between the three choices. In that sense, the thin black lines have a slightly different meaning in the treatments with partner choice compared to the treatments without partner choice.

In combination, the reputation and partner choice mechanisms lead to three treatments besides the Baseline Treatment. The Reputation Treatment implemented only the reputation mechanism, without partner choice. The Partner Choice Treatment, implemented the partner choice mechanism but not the reputation mechanism. Thus, subjects could also choose NO INTERACTION with another subject, but were only informed about what happened in their own interactions.

The Reputation-Partner Choice Treatment implements both the reputation mechanism and the partner choice mechanism. Thus, subjects had the option not to interact (and thus not to be neighbors with another subject), and were informed not only about their own interactions but also about the interactions of their neighbors. It is important to note that in this case, the information network (indicated by the thin black lines in the interface) co-evolves with subjects' behavior in the game: when they choose not to interact with a given other subject, this also has the consequence that they are not (or no longer) informed about the interactions of this other subject. In order to keep this treatment comparable with the Reputation Treatment in terms of information availability, we let subjects start in the complete network in

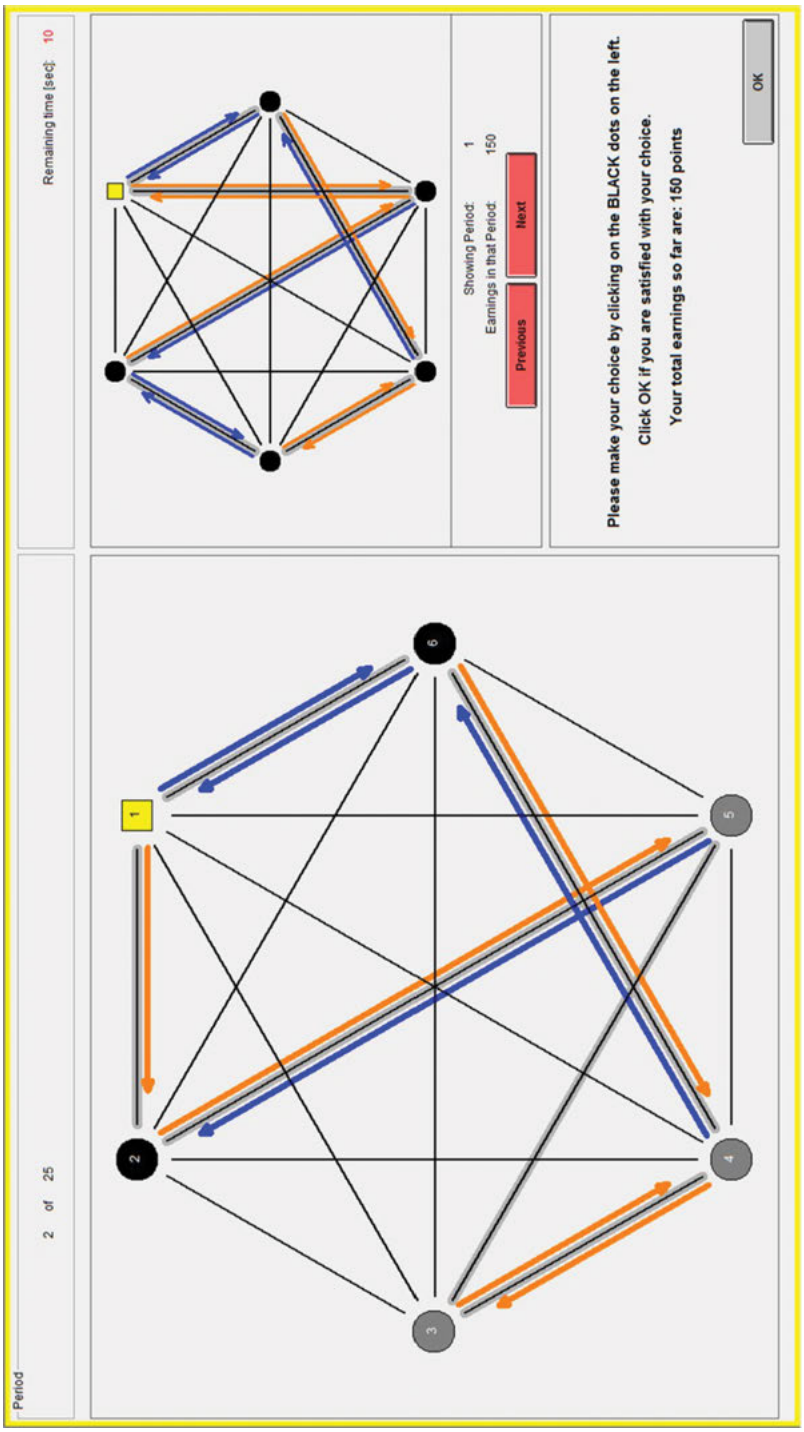


Figure 17.4: The computer interface: the results screen of the Reputation Treatment.

this treatment and neighbors were only disconnected after someone choose NO INTERACTION when they played.[2]

We like to emphasize that an important feature of our design is that all treatments are mutually comparable, apart from the mechanism that we are interested in. Thus, the Reputation Treatment differs from the Baseline Treatment *only* in the implementation of the reputation mechanism, while at the same time (and due to the fact that the no interaction payoff is also implemented in the treatments without partner choice), the Partner Choice Treatment *also* only differ from the Baseline Treatment in the implementation of the partner choice mechanism. Likewise, in the Reputation-Partner Choice Treatment, the implementation of the reputation- and partner choice mechanism is exactly the same as in the corresponding treatments that do not feature the alternate mechanism. Preserving this comparability is not trivial, and it is one of the features that sets our study apart from most of the existing literature.

In all treatments, subjects were instructed about the details of the game and the interface through a set of written instructions, which they had available throughout the experiment for reference. Before the 40 periods of the experiment began, subjects played five "practice periods" to familiarize themselves with the interface and the game. After the 40th period, subjects were shown an overview of the total number of points they had earned. For convenience, we summarize the main elements of the interface:

Thin black lines indicate that subjects both played ORANGE or BLUE the last time they were matched, which implies that they are neighbors and it determines what they observe in the treatments with reputation. The network of neighbors is common knowledge. In the Baseline and Reputation Treatment (treatments without partner choice), all subjects were neighbors.

Thick grey lines indicate that two subjects are matched, which implies that they have the opportunity to choose ORANGE, BLUE or NO INTERACTION in (the treatments with) these interactions. Matches are common knowledge.

Orange and blue arrows indicate choice in the game in the case two subjects are matched, with orange referring to cooperation in the PD and blue to defection. If one of the two subjects chose NO INTERACTION rather than BLUE or ORANGE in the treatments with partner choice, they both received the No interaction payoff.

17.2.4.2 Practical implementation

The experiments were conducted at different moments in time and in two different labs, located at Stanford University and UC Berkeley. Both labs maintain pools of potential

2 Many more example of social dilemma problems and especially in the academic world have vividly been illustrated by Raub in his farewell lecture (Raub 2017: 49–54).

e Treatment in which there is no information exchange about neighbors' interactions with others than Ego. We return to this issue in the discussion.

subjects consisting mostly of undergraduate students, from which subjects were invited to participate. Table 17.1 shows the number of groups in each treatment per location.[3]

Table 17.1: Numbers of groups per treatment and location (N = 55).

	Stanford		Berkeley	
	No Partner Choice	**Partner Choice**	**No Partner Choice**	**Partner Choice**
No Reputation	7	6	6	8
Reputation	7	6	6	9

In each session, either six or twelve subjects participated simultaneously in the lab, resulting in either one or two 6-person groups. The two locations differed substantially with regard to the number of subjects per session: at the Stanford session it happened only rarely that two groups (twelve subjects) were available, while at the Berkeley sessions this was almost always the case. An important consequence of having only one group in the lab was that the subject could not be reshuffled between the practice periods and the actual experiment.

17.3 Results

Figure 17.5 shows the development of average group cooperation rates in each of the treatments. Based on this figure we can make a number of observations. First, we find that the only treatment that clearly stands out is the Partner Choice Treatment (without reputation effects), where cooperation rates are clearly lower than in the other three treatments. Second, we find that in all treatments cooperation rates are fairly stable until approximately period 35, after which we see a mild end-game effect.

The picture becomes somewhat different when we look at the results separately for each of the locations (Figure 17.6). As compared to Figure 17.5, a number of differences stand out. First, cooperation is clearly higher in the Stanford sessions than in the Berkeley sessions. Second, while the Partner Choice Treatment remains the treatment with lowest cooperation rates in both cases, the order of the other three treatments in terms of cooperation rates seems to differ between the locations. In the Stanford sessions, we see a divergence from period 25 onwards between the two treatments with partner choice on the one hand and the two treatments without partner choice on the other. In the Berkeley sessions, the Reputation-Partner Choice Treatment continues to

3 In crosstabs such as Table 17.1, we use combinations of the labels "Reputation"/"No Reputation" and "Partner Choice"/"No Partner Choice" to indicate the experimental treatments, where the cell "No Reputation, No Partner Choice" refers to the Baseline Treatment, and so on.

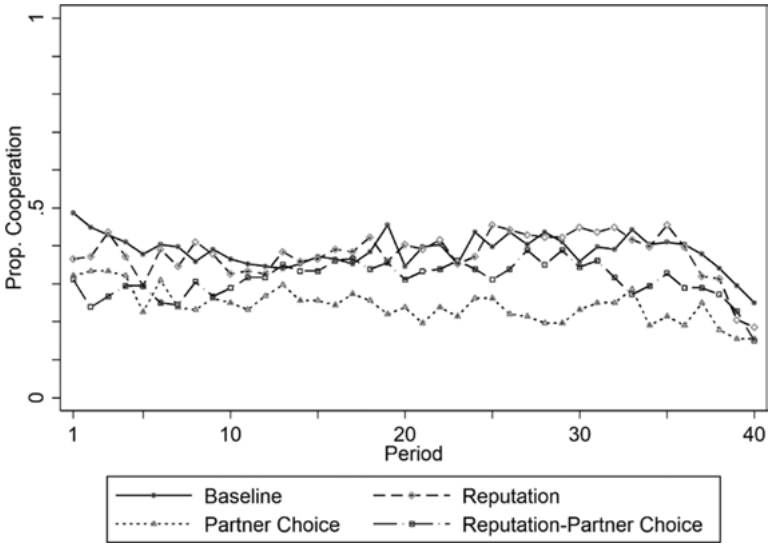


Figure 17.5: Average cooperation per treatment.

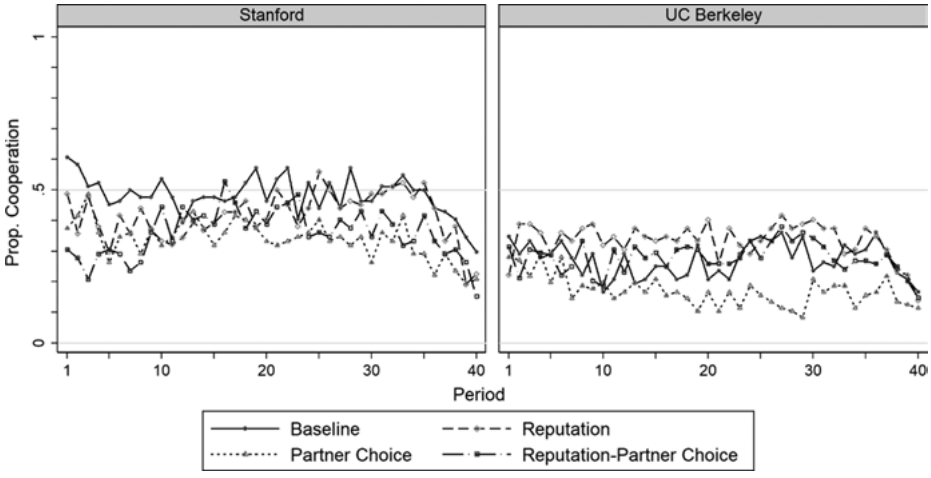


Figure 17.6: Average cooperation rates per treatment, separate for the two locations.

show higher cooperation rates than the Partner Choice Treatment. At the same time, the Reputation Treatment shows consistently (but slightly) higher cooperation rates than the Baseline Treatment. In sum, while we see that the effect of partner choice by itself is consistently negative between the locations, the effect of reputation, both by itself and in combination with partner choice, is less consistent between the locations.

To test these results for statistical significance, we conduct t-tests on average cooperation per treatment, in which the unit of analysis is a group and in which we exclude the final five periods because of the consistently visible endgame effect.[4] We conduct pairwise comparisons of the four treatments to test for effects of partner choice and reputation. The corresponding averages and standard deviations are reported in Tables 17.2 and 17.3.

Table 17.2: Cooperation in periods 1–35 per treatment. Averages and standard deviations.

	No Partner Choice	**Partner Choice**
No Reputation	0.395 (0.135)	0.251 (0.139)
Reputation	0.392 (0.130)	0.320 (0.177)

Table 17.3: Cooperation in periods 1–35 per treatment and location and session size. Averages and standard deviations.

Stanford	**No Partner Choice**	**Partner Choice**
No Reputation	0.498 (0.056)	0.355 (0.118)
Reputation	0.428 (0.155)	0.375 (0.231)
UC Berkeley	**No Partner choice**	**Partner choice**
No Reputation	0.274 (0.088)	0.173 (0.099)
Reputation	0.348 (0.091)	0.285 (0.140)

When we pool the data for the two locations, we find that, as compared to the Baseline Treatment, cooperation is significantly lower in the Partner Choice Treatment ($p = .01$). The other differences between the treatments are not significant.[5] Looking at the locations separately, we find that the Stanford data replicate the overall result: only the difference between the Baseline Treatment and the Partner Choice

4 However, it turns out that the overall results depend very little on the precise choice of periods included.

5 Non-parametric tests yield similar results.

Treatment is significant (p = .02). For the Berkeley data, however, we find that this difference is only significant at the α = .1 level (p = .07), while the difference between the Reputation-Partner Choice Treatment and Partner Choice Treatment is also significant at this level (p = .08). We thus find no support for Hypothesis 1.

To test Hypothesis 2, we run a generalized linear regression model with a logit link (because the depended variable is a proportion) using the implementation in Stata 14. As predictors, we use dummy variables for the treatments (including an interaction between reputation and partner choice), location, and behavior in the first period. Consistent with the hypotheses, behavior in the first period has a strong and significant positive effect on overall cooperation as can be seen in Table 17.4.

Table 17.4: Generalized linear regression of the proportion cooperation on treatments, location, and first period cooperation (N = 55).

	Coeff.	Std. Err.	z
Reputation	0.15	0.18	0.82
Partner Choice	−0.43*	0.20	−2.16
Reputation * Partner Choice	0.25	0.31	0.80
Berkeley	−0.48*	0.18	−2.58
Cooperation Period 1	1.17**	0.43	2.72
Constant	−0.81**	0.30	−2.67

*p < 0.05; **p < 0.01.

In order to test Hypothesis 3, we look at the variance between group cooperation levels. If subjects play trigger strategies, we consider the hypothesized effect to show in later rather than earlier periods of the game. As Table 17.5 shows, the variance in cooperation in periods 30–35 between groups is larger in the Reputation Treatment than in the Baseline Treatment as predicted, but this difference is not significant. The same holds for the Reputation-Partner Choice Treatment and the Partner Choice Treatment: the variance is larger with reputation, but not significantly so. Thus, we

Table 17.5: Cooperation levels in periods 30–35 (means and standard deviations).

	No Partner choice	Partner Choice
No Reputation	0.401	0.237
	(0.155)	(0.165)
Reputation	0.434	0.319
	(0.176)	(0.228)

find no support for Hypothesis 3, and that conclusion does not change if we distinguish between locations (results not shown).

To test Hypothesis 4, we compare cooperation levels in periods 1–35 in *established ties*, that is, interactions in which none of the subjects chose the NO INTERACTION option, if available (see Table 17.6). We hypothesized cooperation in such established ties should be higher in the Partner Choice Treatment than in the Baseline Treatment. As Table 17.6 shows, this is indeed the case, but this difference is not statistically significant ($t = -0.76$, $p = .45$). We also observe that this difference is larger and significant ($t = -2.30$, $p = .03$) if we compare the Reputation Treatment and the Reputation-Partner Choice Treatment, although we did not hypothesize this.

Table 17.6: Cooperation levels in periods 1–35 in established ties, per treatment.

	No Partner Choice	**Partner Choice**
No Reputation	0.395	0.446
	(0.135)	(0.201)
Reputation	0.392	0.553
	(0.130)	(0.220)

Finally, we briefly explore emerging network structure in the two partner choice treatments by looking at the dynamics of network density, where the network is defined as the network of interactions in which no subject in the pair chose exit. In the two reputation treatments, this network determined the flow of information (see "Experimental setup"). Figure 17.7 shows the development of network density for each of the two treatments. Both treatments start in the complete network (although in the treatment without reputation this network is not used to transfer information)[6] but in both cases network density quickly decreases until approximately period 12, after which density only decreases slightly.[7] In combination with our earlier results on the dynamics of cooperation (Figure 17.5), which showed more constant or even slightly increasing cooperation levels, the decrease in density indicates that especially interactions involving defection disappear. After period 12, density in the

6 Note that this does not necessarily imply that both subjects played C or D; in the earlier periods, it is possible that subjects had not yet had an opportunity to change the tie, as they could only change two ties per person in each period. For comparability between the treatments with and without reputation, we treat these ties as present in the density calculation for Figure 17.7.

7 Because we measured the network density at the end of each period, the average densities in the first period are not necessarily one, as subjects could already change the network during that period.

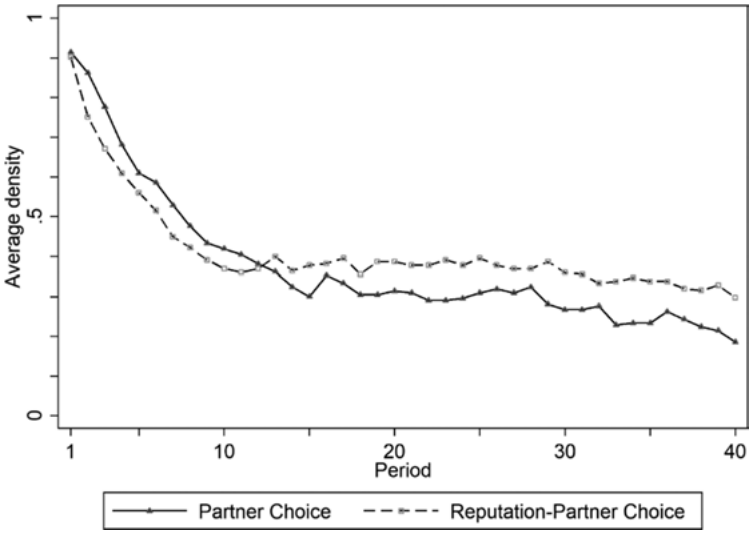


Figure 17.7: Dynamics of network density for each of the partner choice treatments.

Reputation-Partner Choice Treatment appears slightly higher than in the Partner Choice Treatment, but this difference is not statistically significant (taking period 35 as a benchmark, consistent with our earlier analyses of emerging cooperation levels).

17.4 Discussion and conclusion

The notion that cooperation is more like in more cohesive communities is almost a truism in sociological thinking. Nevertheless, providing a rigorous theoretical argument for this claim is far from trivial, as Raub and Weesie (1990) showed in their landmark paper. In this paper, we have extended their scenario of repeated Prisoner's Dilemmas in social networks with the option of partner choice. This option makes social network structure effectively endogenous, and allows us to study the question whether the social network structures that are thought to promote cooperation (i.e., dense network structures) are likely to remain or emerge as a result of actor's choices, or whether the option of partner choice undermines the reputation effects in networks theorized by Raub and Weesie.

Characteristic for our experiment is that the experimental treatments are specifically designed to 1) test the prediction from the Raub and Weesie (1990) model and 2) to study the effects of partner choice and reputation diffusion in isolation as well as in combination. After all, previous research suggests that the option of partner choice in itself affects cooperation rates, even in the absence of network effects (e.g., Hauk 2001). Thus, our experimental treatments are constructed such that

both partner choice and network effects via reputation diffusion are varied independently while keeping other parameters constant. This approach results in four experimental treatments: the Baseline Treatment without reputation effects or partner choice, the Partner Choice Treatment with only partner choice, the Reputation Treatment with only information exchange, and the combined Reputation-Partner Choice Treatment. Another element that distinguishes our experiment from most other experiments in which subjects play PDs on networks is that subjects can differentiate their behavior between different others they play with. This reduces some of the interdependencies between different interactions, but it is crucial for allowing targeting punishment to specific individuals by defecting in the interactions with these individuals and being able to cooperate with others.

Based on the work by Raub and Weesie (1990), we hypothesized in the first place that cooperation rates should be higher in the Reputation Treatment than in the Baseline Treatment (H1). As we already reported in an earlier paper (Corten et al. 2016), we found no support for this prediction, whether we look at overall cooperation rates, cooperation rates near the end of the game, or cooperation rates at the start of the game (the last is arguably the most precise test of the hypothesis, see Corten et al. 2016). Alternative models of cooperation that rely on learning heuristics rather than forward-looking rationality (i.e., Corten and Cook 2009) predict that emerging cooperation levels (near the end of the game) should be higher when initial cooperation levels are higher (H2). We do find support for this hypothesis, but we do not find support for the additional hypothesis, predicted by the same models, that the variance in cooperation is higher in the Reputation Treatment than in the Baseline Treatment (H3).

For the Partner Choice Treatment, we argued that the availability of a relatively attractive outside option would reduce overall cooperation rates as it reduces the possibility to sanction defection effectively (cf. Hauk and Nagel 2001; Hauk 2001), while increasing cooperation rates *conditional* on that both actors enter into the interaction (H4), since actors should only engage in interaction if they believe that the other actor will cooperate. While we do not find support for the latter, we do find that overall cooperation rates are lower in the Partner Choice Treatment than in the Baseline Treatment, which is in fact the only treatment-level difference in cooperation rates that is statistically significant. Thus, while we find that the option of partner choice lowers cooperation rates in the absence of reputation effects, this is not the case in the presence of reputation effects (that is, there are no significant differences in cooperation rates between the Reputation Treatment and the Reputation-Partner Choice Treatment). Although it is tempting to interpret this finding as showing that the presence of reputation effects serves as "protection" against the detrimental effect of partner choice, we note that there is no significant negative interaction effect between the Partner Choice- and Reputation Treatments (as this interpretation would imply).

Looking at the emergence of network structure in the two treatments with partner choice, we did not find clear differences in network structures between the treatments with and without reputation: average densities converge to about .3, which is obviously

considerably less dense than the complete network in the reputation treatment without partner choice. What does this result tell us about the likelihood of the spontaneous emergence of network structures that support cooperation? On the one hand, we do not find that if subjects can endogenously change the network, this leads to high-density networks, nor do we find that the mere availability of a reputation mechanism, which could be interpreted as an incentive for striving for dense networks, clearly leads to differences in density of emerging network structures. On the other hand, we do also not find that reputation formation in dense networks, when determined exogenously, leads to more cooperation in the first place, which renders the question of emergence of cooperation-promoting network structure less relevant altogether, in the context of our experiment. Similar to what is found in the context of trust games (Frey et al. 2019), we cannot exclude that subjects do not stay in dense networks because they do not anticipate the potential benefits or that they do not stay in these networks because they do anticipate that the potential benefits will not materialize.

Our setup did not allow us to study the emergence of social networks (with information transmission) in the case where actors start completely isolated: in our Reputation-Partner Choice Treatment, subjects already started in the completely connected network to ensure comparability with the (static) Reputation Treatment. On the one hand, studying such an extension (e.g., a version of the Reputation-Partner Choice Treatment in which subjects start in the empty network) may provide novel insights in the types of network structures that emerge when subjects are free to build the network structure "from scratch." On the other hand, given our current findings, there seems little reason to expect that this would lead to different conclusions with regard to cooperation levels.

Before we move on to implications for further research, we also note that we observed considerable differences, both in overall cooperation rates and in the effects of the experimental treatments, between the two location where the experimental sessions took place. Our data do not allow us to disentangle group size effects, differences in the compositions of the respective subject pools, or differences in the setup of the labs underlying these discrepancies, but at the very least these findings illustrate the sensitiveness of experimental results on cooperation issues, even in an incentivized, neutrally worded experiment using the exact same protocol in two locations in the same geographical region. One explanation about these differences might be related to the different populations in the different locations. As Melamed et al. (2017) show, cooperation is predominantly promoted through dynamic networks if the actors involved are sufficiently pro-social or, alternatively, as is suggested by Buskens et al. (2015), subjects differ in the extent to which they anticipate the future benefits of reputation effects. Delving deeper into differences in effects of social structures on cooperation depending on the individual differences between the actors involved is still an important area for further research.

Clearly there are also other directions for extension that are theoretically and empirically interesting. First, the role of noise in the observation of behavior is an

important aspect that is also relevant in real-world context and might intervene in the dynamics of networks of cooperation especially when reputation effects are important. Some work in this direction has been done (e.g. Perc 2006), but many pathways are also unexplored. Another interesting extension is a context in which some relations are dynamic, but other relations cannot be changed. An interesting pioneering study for such a context shows that cooperative behavior in the two types of ties can be interrelated (Harrell, Melamed, and Simpson 2018).

References

Axelrod, Robert, and William D. Hamilton. 1981. "The Evolution of Cooperation." *Science* 211, no. 4489: 1390–1396.

Axelrod, Robert. 1981. "The Emergence of Cooperation among Egoists." *American Political Science Review* 75, no. 2: 306–318.

Bowles, Samuel, and Herbert Gintis. 2013. *A Cooperative Species: Human Reciprocity and Its Evolution*. Princeton, NJ: Princeton University Press.

Boone, Thomas R., and Michael W. Macy. 1999. "Unlocking the Doors of the Prisoner's Dilemma: Dependence, Selectivity, and Cooperation." *Social Psychology Quarterly* 62, no 1: 32–52.

Burt, Ronald S., and Marc Knez. 1995. "Kinds of Third-Party Effects on Trust." *Rationality and Society* 7, no. 3: 255–292.

Buskens, Vincent. 2002. *Social Networks and Trust*. Dordrecht: Kluwer Academic Publishers.

Buskens, Vincent, Werner Raub, Nynke van Miltenburg, Estrella R. Montoya, and Jack van Honk. 2016. "Testosterone Administration Moderates Effect of Social Environment on Trust in Women Depending on Second-to-Fourth Digit Ratio." *Scientific Reports* 6: 27655.

Casella, Alessandra, Navin Kartik, Luis Sanchez, and Sébastien Turban. 2018. "Communication in Context: Interpreting Promises in an Experiment on Competition and Trust." *Proceedings of the National Academy of Sciences* 115, no. 5: 933–938.

Cassar, Alessandra. 2007. "Coordination and Cooperation in Local, Random and Small World Networks: Experimental Evidence." *Games and Economic Behavior* 58, no. 2: 209–230.

Coleman, James S. 1964. *Introduction to Mathematical Sociology*. New York: The Free Press.

Coleman, James S. 1986. "Social Theory, Social Research, and a Theory of Action." *American Journal of Sociology* 91, no. 6: 1309–1335.

Coleman, James S. 1990. "Norm-Generating Structures." Pp. 250–273 in *The Limits of Rationality*, eds. Karen S. Cook and Margaret Levi. Chicago, IL: University of Chicago Press.

Corten, Rense, and Karen S. Cook. 2009 "Cooperation and Reputation in Dynamic Networks." In *Proceedings of the First International Conference on Reputation: Theory and Technology – ICORE*, vol. 9, pp. 20–34.

Corten, Rense, and Vincent Buskens. 2010. "Co-evolution of Conventions and Networks: An Experimental Study." *Social Networks* 32, no. 1: 4–15.

Corten, Rense. 2014. *Computational Approaches to Studying the Co-evolution of Networks and Behavior in Social Dilemmas*. Chichester: Wiley & Sons.

Corten, Rense, Stephanie Rosenkranz, Vincent Buskens, and Karen S. Cook. 2016. "Reputation Effects in Social Networks Do Not Promote Cooperation: An Experimental Test of the Raub & Weesie Model." *PloS One* 11, no. 7: e0155703.

Corbae, Dean, and John Duffy. 2008. "Experiments with Network Formation." *Games and Economic Behavior* 64, no. 1: 81–120.

Cuesta, Jose A., Carlos Gracia-Lázaro, Yamir Moreno, and Angel Sánchez. 2018. "Reputation is Required for Cooperation to Emerge in Dynamic Networks." *arXiv preprint arXiv:1803.06035*.

The EdK-Group. 2000. "Exit, Anonymity and the Changes of Egoistical Cooperation." *Analyse und Kritik* 22, no.1: 114–129.

Ellickson, Robert. C. 1991. *Order Without Law. How Neighbors Settle Disputes* Boston, MA: Harvard University Press.

Falk, Armin, and James J. Heckman. 2009. "Lab Experiments Are a Major Source of Knowledge in the Social Sciences." *Science* 326, no. 5952: 535–538.

Fehl, Katrin, Daniel J. van der Post, and Dirk Semmann. 2011. "Co-evolution of Behaviour and Social Network Structure Promotes Human Cooperation." *Ecology Letters* 14, no. 6: 546–551.

Fehr, Ernst, and Simon Gächter. 2002. "Altruistic Punishment in Humans." *Nature* 415, no. 6868: 137.

Fischbacher, Urs. 2007. "z-Tree: Zurich Toolbox for Ready-Made Economic Experiments." *Experimental Economics* 10, no. 2: 171–178.

Frey, Vincenz, Vincent Buskens, and Rense Corten. 2019. "Investments in and Returns on Network Embeddedness: An Experiment with Trust Games." *Social Networks* 56: 81–92.

Fu, Feng, Christoph Hauert, Martin A. Nowak, and Long Wang. 2008. "Reputation-Based Partner Choice Promotes Cooperation in Social Networks." *Physical Review E* 78, no. 2: 026117.

Gracia-Lázaro, Carlos, Alfredo Ferrer, Gonzalo Ruiz, Alfonso Tarancón, José A. Cuesta, Angel Sánchez, and Yamir Moreno. 2012. "Heterogeneous Networks Do Not Promote Cooperation When Humans Play a Prisoner's Dilemma." *Proceedings of the National Academy of Sciences* 109, no. 32: 12922–12926.

Gracia-Lázaro, Carlos, José A. Cuesta, Angel Sánchez, and Yamir Moreno. 2012. "Human Behavior in Prisoner's Dilemma Experiments Suppresses Network Reciprocity." *Scientific Reports* 2: 325.

Granovetter, Mark. 1985. "Economic Action and Social Structure: The Problem of Embeddedness." *American Journal of Sociology* 91, no. 3: 481–510.

Greif, Avner. 1989. "Reputation and Coalitions in Medieval Trade: Evidence on the Maghribi Traders." *The Journal of Economic History* 49, no. 4: 857–882.

Greif, Avner. 1994. "Cultural Beliefs and the Organization of Society: A Historical and Theoretical Reflection on Collectivist and Individualist Societies." *Journal of Political Economy* 102, no. 5: 912–950.

Grujić, Jelena, Constanza Fosco, Lourdes Araujo, José A. Cuesta, and Angel Sánchez. 2010. "Social Experiments in the Mesoscale: Humans Playing a Spatial Prisoner's Dilemma." *PloS One* 5, no. 11: e13749.

Harrell, Ashley, David Melamed, and Brent Simpson. 2018. "The Strength of Dynamic Ties: The Ability to Alter Some Ties Promotes Cooperation in Those That Cannot Be Altered." *Science Advances* 4, no. 12: eaau9109.

Hauk, Esther. 2001. "Leaving the Prison: Permitting Partner Choice and Refusal in Prisoner's Dilemma Games." *Computational Economics* 18, no. 1: 65–87.

Hauk, Esther, and Rosemarie Nagel. 2001. "Choice of Partners in Multiple Two-Person Prisoner's Dilemma Games: An Experimental Study." *Journal of Conflict Resolution* 45, no. 6: 770–793.

Homans, George C. 1958. "Social Behavior as Exchange." *American Journal of Sociology* 63, no. 6: 597–606.

Kirchkamp, Oliver, and Rosemarie Nagel. 2007. "Naive Learning and Cooperation in Network Experiments." *Games and Economic Behavior* 58, no. 2: 269–292.

Macaulay, Stewart. 1963. "The Use and Nonuse of Contracts in the Manufacturing Industry." *Practical Lawyer* 9, no. 7: 13–40.

Melamed, David, and Brent Simpson. 2016. "Strong Ties Promote the Evolution of Cooperation in Dynamic Networks." *Social Networks* 45: 32–44.
Melamed, David, Ashley Harrell, and Brent Simpson. 2018. "Cooperation, Clustering, and Assortative Mixing in Dynamic Networks." *Proceedings of the National Academy of Sciences* 115, no. 5: 951–956.
Melamed, David, Brent Simpson, and Ashley Harrell. 2017. "Prosocial Orientation Alters Network Dynamics and Fosters Cooperation." *Scientific Reports* 7, no. 1: 357.
Orbell, John M., and Robyn M. Dawes. 1993. "Social Welfare, Cooperators' Advantage, and the Option of not Playing the Game." *American Sociological Review* 58, no. 6: 787–800.
Ostrom, Elinor. 1990. *Governing the Commons: The Evolution of Institutions for Collective Action*. Cambridge: Cambridge University Press.
Parsons, Talcott. 1937. *The Structure of Social Action, Vol. 1*. New York: Free Press.
Pennisi, Elizabeth. 2005. "How Did Cooperative Behavior Evolve?" *Science* 309, no. 5731: 93–93.
Perc, Matjaž. 2006. "Double Resonance in Cooperation Induced by Noise and Network Variation for an Evolutionary Prisoner's Dilemma." *New Journal of Physics* 8, no. 9: 183.
Rand, David G., Samuel Arbesman, and Nicholas A. Christakis. 2011. "Dynamic social networks promote cooperation in experiments with humans." *Proceedings of the National Academy of Sciences* 108, no. 48: 19193–19198.
Rand, David G., Martin A. Nowak, James H. Fowler, and Nicolas A. Christakis. 2014. "Static Network Structure Can Stabilize Human Cooperation." *Proceedings of the National Academy of Sciences*, no. 111: 17093–17098.
Raub, Werner. 2017. *Rational Models*. Utrecht: Utrecht University.
Raub, Werner. 1984. *Rationale Akteure, institutionelle Regelungen und Interdependenzen: Untersuchungen zu einer erklärenden Soziologie auf strukturell-individualistischer Grundlage*. Frankfurt a.M.: Lang.
Raub, Werner. 1997. *Samenwerking in Duurzame Relaties en Sociale Cohesie (Cooperation in Durable Relations and Social Cohesion)*. Inaugural Lecture. Utrecht: Utrecht University.
Raub, Werner, and Vincent Buskens. 2008. "Theory and Empirical Research in Analytical Sociology: The Case of Cooperation in Problematic Social Situations." *Analyse & Kritik* 30, no. 2: 689–722.
Raub, Werner, Vincent Buskens, and Vincenz Frey. 2013. "The Rationality of Social Structure: Cooperation in Social Dilemmas Through Investments in and Returns on Social Capital." *Social Networks* 35, no. 4: 720–732.
Raub, Werner, Vincent Buskens, and Rense Corten. 2015. "Social Dilemmas and Cooperation." Pp. 597–626 in *Handbuch Modellbildung und Simulation in den Sozialwissenschaften*, eds. Norman Braun and Nicole J. Saam. Wiesbaden: Springer VS.
Raub, Werner, Vincent Buskens, and Marcel A. L. M. van Assen. 2011. "Micro-Macro Links and Microfoundations in Sociology." *Journal of Mathematical Sociology* 35, no. 1–3: 1–25.
Raub, Werner, and Thomas Voss. 1981. *Individuelles Handeln und gesellschaftliche Folgen: das individualistische Programm in den Sozialwissenschaften*. Darmstadt: Luchterhand.
Raub, Werner, and Thomas Voss. 1986a."Conditions for Cooperation in Problematic Social Situations." Pp. 85–103 in *Paradoxical Effects of Social Behavior*, eds. Andreas Diekmann and Peter Mitter. Heidelberg: Physica-Verlag.
Raub, Werner, and Thomas Voss. 1986b. "Die Sozialstruktur der Kooperation rationaler Egoisten." *Zeitschrift für Soziologie* 15, no. 5: 309–323.
Raub, Werner, and Jeroen Weesie. 1990. "Reputation and Efficiency in Social Interactions: An Example of Network Effects." *American Journal of Sociology* 96, no. 3: 626–654.

Raub, Werner, and Jeroen Weesie. 2000. "The Management of Matches: A Research Program on Solidarity in Durable Social Relations." *The Netherlands Journal of Social Sciences* 36, no. 1: 71–88.

Riedl, Arno, and Aljaz Ule. 2002. "Exclusion and Cooperation in Social Network Experiments." Unpublished Manuscript, Amsterdam: University of Amsterdam.

Rooks, Gerrit, Werner Raub, Robert Selten, and Frits Tazelaar. 2000. "How Inter-Firm Co-operation Depends on Social Embeddedness: A Vignette Study." *Acta Sociologica* 43, no. 2: 123–137.

Sánchez, Angel. 2018. "Physics of Human Cooperation: Experimental Evidence and Theoretical Models." *Journal of Statistical Mechanics: Theory and Experiment* 2018, no. 2: 024001.

Schuessler, Rudolf. 1989. "Exit Threats and Cooperation under Anonymity." *Journal of Conflict Resolution* 33, no. 4: 728–749.

Stanley, E. Ann, Dan Ashlock, and Mark D. Smucker. 1995. "Iterated Prisoner's Dilemma with Choice and Refusal of Partners: Evolutionary Results." In *European Conference on Artificial Life*, 490–502. Berlin and Heidelberg: Springer.

Taylor, Michael. 1976."Anarchy and Cooperation." *Political Theory* 5, no. 2: 271–275.

Ule, Aljaž. 2005. "Exclusion and Cooperation in Networks." PhD thesis, Tinbergen Institute.

Uzzi, Brian. 1996. "The Sources and Consequences of Embeddedness for the Economic Performance of Organizations: The Network Effect." *American Sociological Review*: 61, no 4: 674–698.

Uzzi, Brian. 1997. "Social Structure and Competition in Interfirm Networks: The Paradox of Embeddedness." *Administrative Science Quarterly* 42, no. 1: 35–67.

Voss, Thomas. 1982. "Rational Actors and Social Institutions: The Case of the Organic Emergence of Norms." Pp. 76–100 in *Theoretical models and empirical analyses: Contributions to the explanation of individual actions and collective phenomena*, ed. Werner Raub. Utrecht: ESP.

Voss, Thomas. 2001. "Game-Theoretical Perspectives on the Emergence of Social Norms." Pp. 105–136 in *Social Norms*, eds. Michael Hechter and Karl-Dieter Opp. New York: Russell Sage.

Wilson, Alistair J., and Hong Wu. 2017. "At-Will Relationships: How an Option to Walk Away Affects Cooperation and Efficiency." *Games and Economic Behavior* 102: 487–507.

Davide Barrera, Vincent Buskens and Vera de Rover

18 Comparing Consequences of Carrots and Sticks on Cooperation in Repeated Public Good Games

Abstract: Many sociologists and economists have maintained that costly sanctions are able to sustain cooperation, but whether carrots or sticks are more successful in this respect is still under dispute (e.g., Balliet, Mulder, and Van Lange 2011; Rand et al. 2009; Sefton, Schupp, and Walker 2007). Furthermore, while many studies investigated the effects of sanctioning institution on cooperation, the long-term effects of sanctions on group solidarity are largely unexplored. In this chapter, we discuss contrasting hypotheses concerning the effects of positive and negative sanctions on cooperation in Public Good Games and solidarity among the group members. Subsequently, we test these hypotheses by means of a laboratory experiment. Our results show that while carrots do increase cooperation, sticks turn out to be more effective. Concerning group solidarity, we do not find differences in group solidarity depending on the type of sanctions available to the group members. However, we find that actors who receive rewards show higher solidarity towards the group.

18.1 Introduction

In many practical instances, individual interests are partly conflicting with collective interests. For example, everyone prefers to receive public services from the state, but, at the same time, everyone would prefer not to pay taxes that allow the state to provide those services. As it is individually rational to avoid paying taxes, but collectively detrimental if everyone does so, we call this a social dilemma (Raub, Buskens, and Corten 2015). Real-life situations with this kind of incentive structure are modelled using public good games (PGG hereafter). The typical setup of a PGG is as follows. People interact in small groups. Each individual receives an initial endowment and can decide how much of this endowment he or she wants to invest into a 'group account,' and how much he or she wants to keep for him- or herself. In a PGG, the amount contributed by all group members to the group account grows by a certain rate and – thereafter – is equally distributed among all the group members, regardless of their contribution to the public good. The growth rate of the public good is always lower than the number of people in the group. This

Davide Barrera, University of Turin and Collegio Carlo Alberto
Vincent Buskens, Vera de Rover, Department of Sociology / ICS, Utrecht University

https://doi.org/10.1515/9783110647495-018

implies that, for every unit an actor invests in the public good, he or she receives only a fraction in return. Therefore, it is individually rational not to invest in the group account.

Although standard game-theoretical rationality assumptions dictate that nobody should cooperate in public good games, some cooperation is always observed when subjects play public good games in experimental laboratories (see Chauduri 2011 and Ledyard 1995 for reviews). Several mechanisms accounting for the emergence of cooperation have been proposed, such as kin selection (Hamilton 1963), direct reciprocity (Trivers 1971), social embeddedness (Buskens and Raub 2013; Raub and Weesie 1990), indirect reciprocity (Nowak and Sigmund 1998), group selection (Traulsen and Nowak 2006), status (Willer 2009). See Nowak (2006) for an overview. In the context of cooperation in PGGs, a large body of literature focused on the role of institutions. Analyzing the problem of social order, Hobbes (1651) described an authority who enforces rules for the society by establishing a contract. Social order obtains because violations of this contract are punished by the central authority (the "Leviathan" in Hobbes's terms). More generally, institutional solutions to public good problems include all forms of sanctions that can be positive (i.e., rewards or carrots) or negative (i.e., punishments or sticks).[1] If actors are given the possibility to use sanctions, they typically use them as selective incentives to promote cooperation (Olson 1965). The idea that sanctions are an effective means to protect cooperation from free-riders is widely known and has been largely debated (Balliet, Mulder, and van Lange 2011; Barclay 2006; Fehr and Gächter 2002; Fehr and Gintis 2007; Gächter 2013; Yamagishi 1986). While Balliet, Mulder, and Van Lange (2011) convincingly show – using an extensive meta-analysis – that both sticks and carrots can promote cooperation and that the two effects do not seem too much different, the number of studies with costly rewards in PGGs is limited. In addition, the consequences of using sanctions, particularly punishment, have been considered only in terms of their ability to promote cooperation. Yet, the fact that Hobbes (1651) used a biblical monster, the leviathan, as a metaphor for the central authority responsible of maintaining social order, suggests that order imposed by force is not necessarily always pleasant. In fact, some scholars have addressed possible side-effects of punitive sanctions concerning solidarity among the groups members (Fehr and Rockenbach 2003; Irwin, Mulder, and Simpson 2014; Mulder et al. 2006; Nowak et al. 2008; Oliver 1980). In this chapter, we focus on two questions: first, whether carrots or sticks are a more effective way to obtain cooperation in a public good problem, and second, what are the consequences of positive and negative sanctions on how much solidarity individuals feel towards their fellow group members. In the remainder, after briefly discussing the existing relevant literature, we present a series of hypotheses concerning effects of positive and negative

1 See Calvert (1995) on modeling institutions as exogenous constraints.

sanctions on both cooperation and solidarity. Subsequently, we present an experiment in which these hypotheses are empirically tested.

18.2 Theory and hypotheses

18.2.1 Theory: Human cooperation and selective incentives

Before discussing the theoretical arguments related to the use of positive versus negative sanctions, we formally introduce the PGG. In a PGG, groups consist of N (with $N > 2$) persons. Each group member is endowed with a specific amount (Y) of points. The amount of points is the same for every person. Each member can invest a certain amount X ($0 \leq X \leq Y$) of these points into a group account representing the public good. Every point invested in the group account is multiplied by a factor M (which exceeds 1 but is smaller than N, $1 < M < N$). After all group members have made their contribution to the group account and the total amount of points has been multiplied by M, the resulting amount of points is equally distributed among the group members. Because $M < N$, the individual return of the investment is negative, i.e. $X > X \times M/N$. Therefore, choosing $X = 0$ is a dominant strategy for every player. Thus, assuming that actors are purely interested in maximizing their own wealth, they should not contribute anything to the public good. However, if all group members invest their entire endowment Y, each individual earns $N \times Y \times M/N = X \times Y > Y$. This situation is Pareto-superior to the situation in which no one invests anything (Fehr and Gintis 2007; Gächter 2013). In a one-shot PGG, if standard economic rationality is assumed (that is, everyone tries to maximize his expected payoff with no concern for what happens to the other group members), all actors should play the dominant strategy and choose to contribute nothing to the public good. However, this hypothesis is typically not supported by experimental results. In the laboratory there are always some persons who cooperate: in one-shot PGGs, the average contribution level is usually around 50% of the total endowment.

If the PGG is played repeatedly, simple reciprocity (Hamilton 1963; Trivers 1971) signaling (Spence 1973; Zahavi 1995) or reputation (Axelrod 1984; Kreps et al. 1982; Raub and Weesie 1990) become possible mechanisms supporting contribution to the public good. Nevertheless, repeated PGGs typically show a pattern of decreasing contribution over time empirically. In experimental treatments in which the group composition is kept constant (*partner* treatment), the mean contribution is higher than when participants are shuffled after every round (*stranger* treatment), but in both cases contribution declines as the game proceeds further (e.g., Andreoni 1988; Keser and Van Winden 2000). These empirical results imply that reciprocity, reputation and costly signaling are not sufficient to reach a cooperative equilibrium in a repeated PGG. The pattern of declining cooperation is reversed and higher cooperation

levels are achieved and sustained when sanctions are allowed (Fehr and Gächter 2002) – even if participants are shuffled and re-matching is excluded by design. Fehr and Gachter's (2002) study on altruistic punishment shows that 1) a significant proportion of actors has not purely selfish preferences, 2) these actors typically contribute significantly to the public good and punish defectors, 3) as defectors anticipate that they will be punished, they raise their cooperation levels, and, consequently, cooperation is ensured. The existence of (partly) altruistic preferences poses an evolutionary puzzle because selfish actors gain a competitive advantage in interactions with these altruists. Therefore, altruistic preferences should have been driven extinct by natural selection.

Some scholars have argued that these preferences are a form of maladaption (e.g., Johnson, Stopka, and Knights 2003). The argument goes as follows: in the ancestral environment where most of our evolution took place – sometimes referred to as "environment of evolutionary adaptiveness", EEA (e.g., Fehr and Henrich 2003) – interactions were typically *repeated* rather than one-shot, because most of the relevant social activities occurred within small groups. In such settings, reputation-building makes cooperation profitable, because defection might lead to retaliation and conflict, thus making defectors worse off in the long run. As one-shot interactions are assumed to have been rare in EEA, human subjects interacting anonymously in the laboratory fail to recognize the irrelevance of reputation building in one-shot interactions and erroneously apply behavioral rules evolved and adapted to repeated interactions.

However, this argument is problematic because one-shot interactions were in fact not so rare in EEA and, moreover, experimental subjects do show the ability to differentiate their behavior depending on whether interactions are repeated or one-shot (Fehr and Henrich, 2003). To solve this puzzle, scholars have developed culture-gene coevolution models that provide evolutionary foundations for altruistic preferences (e.g., Boyd et al. 2003). In such models actors face cooperation problems within groups and, simultaneously, competition for scarce resources between groups. Consequently, selection favors defectors within groups, but groups where defectors thrive underperform in intergroup competition and risk extinction. In this environment, groups developed some form of institutional solution – such as punishment or reward – to prevent defectors to exploit cooperators thereby producing a reduction of the group collective fitness, ultimately leading to extinction of the group. Thus, these models typically include some form of group selection, but they assume that group selection operates predominantly on cultural rather than genetic variation (Richerson, Boyd and Henrich 2003). Accordingly, altruistic preferences evolved in the EEA lead to the development of various institutions, which in turn are transmitted to subsequent generations through processes like conformist social learning and moralistic enforcement of norms (Richerson, Boyd and Henrich 2003). However, if institutions are assumed to be transmitted via conformist social learning, group solidarity and individual commitment to the group must be important determinants of a group's capacity to survive and

reproduce itself. Yet, the relation between sanctioning institutions and group solidarity has been largely ignored.

In this chapter we study a setting in which both positive and negative sanctions are allowed and we focus on a repeated game with the same partners because earlier results revealed that, in these conditions, a strong effect of sanctioning institutions is to be expected (Balliet, Mulder, and van Lange 2011; Rand et al. 2009; Sefton, Schupp, and Walker 2007). Also, we focus on groups that remain stable over time, because we are interested in possible side-effects of sanctions, particularly with respect to group solidarity, which only make sense if the group members experience more than just one interaction with each other.

18.2.2 Punishment and rewards

While reciprocity can produce cooperation in two-person games (Axelrod 1984), when the social dilemma involves a group of actors, as in the PGG, reciprocity alone is insufficient, because the individual actions are not *selective*. That is, when the payoff of an actor is affected by more than one person, it is not possible to apply direct reciprocity. In a PGG, either one cooperates with all other group members or with none. Therefore, cooperation is unstable, even in repeated games with the same partners (Andreoni 1988; Keser and Van Winden 2000). The option to impose sanctions, whether in the form of carrots (rewards) or sticks (punishments) addresses precisely this problem of *selecting* incentives to discriminate cooperators from free-riders. By being able to address only the defectors with sticks, or only the contributors with carrots, cooperation can possibly be sustained (Andreoni, Harbaugh, and Vesterlund 2003; Barclay 2006; Fehr and Gächter 2002; Masclet et al. 2003; Rand et al. 2009; Sefton, Schupp, and Walker 2007; Stoop, Van Soest, and Vyrastekova 2018; Yamagishi 1986).

Sanctions are common in everyday life: even just looking at someone angrily or smiling at someone in response to some action could be seen as a stick or carrot, respectively (cf. Masclet et al. 2003; Pan and Houser 2017; Takacs and Janky 2007). In experimental settings, punitive monetary sanctions are often implemented by allowing everyone to impose a (costly) fine on anyone else. For example, each member of a specific group can decrease or increase each other group members' payoffs by paying a small amount from their own earnings. Similarly, rewards can be implemented by allowing everyone to increase the payoff of anybody else, paying a small cost.

According to standard game theoretical assumptions, selective sanctioning institutions imply an additional problem as they likewise require cooperation to be produced. This problem is commonly known as the second-order free-rider dilemma (Yamagishi 1986). Although theoretically rational self-interested actors should not use costly sanctions, many experiments have shown that, when given the opportunity, a considerable proportion of actors invest resources to punish defectors or reward

cooperators. In order to account for this empirical evidence, alternative assumptions concerning the preferences of the actors have been proposed. e.g., *strong reciprocity* (Fehr and Gintis 2007; Gintis 2000). Individuals are assumed to be motivated by strong reciprocity if they cooperate when they expect others to cooperate, reciprocate when they are rewarded and punish violations of the cooperative norm (Fehr and Gintis 2007; Keser and Van Winden 2000). Most commonly, cooperators (people who invest) punish defectors (people who do not invest into the group account) and reward other cooperators (Fehr and Gächter 2002; Gächter 2013; Sefton, Schupp, and Walker 2007).[2] Consequently, as defectors anticipate that they will be punished or that they will not receive any reward, they raise their contribution levels (Fehr and Gächter 2002). Accordingly, the classical model of economic rationality – *homo economicus* – does not accurately account for human behavior in real life. An alternative model, incorporating assumptions of strong reciprocity, is commonly referred to as the PBC-model (Preferences, Beliefs and Constraints- model) (Fehr and Gintis 2007; Gächter 2013). This model considers *P*references, *B*eliefs and *C*onstraints, on the basis of which actors make their decisions. "*Preferences* describe how an individual ranks the available alternatives according to his or her subjective tastes" (Gächter 2013: 34) and these can include more aspects than just his or her own money. That is, partly altruistic preferences are also possible. *Beliefs* refer to what actors think the preferences of other actors might be or how these preferences are distributed. Finally, *Constraints* "describe the set of alternatives that are available to an individual" (Gächter 2013: 34). The PBC-model assumes that individuals choose their utility-maximizing option, an option that satisfies their preferences best, while considering their beliefs and constraints.

18.2.3 Hypotheses on contributions to the public good

In order to derive hypotheses on the effects of monetary sanctions (carrots and sticks) on cooperation in public good type of settings, we assume that actors make their decisions as postulated by the PBC-model. The PBC-model can accommodate various types of preferences, including inequality aversion (Fehr and Schmidt 1999) or fairness (Rabin 1993). Although distinguishing between these alternative non-standard utility models exceeds the scope of our chapter, preferences for fairness and equality are important motivational forces behind the phenomenon of strong reciprocity (Fehr and Gintis 2007; Gächter 2013; Keser and Van Winden 2000). Consistent with the PBC-model, we distinguish two types of individuals: *selfish individuals* (that correspond to the homo economicus) and *conditional cooperators*

2 Antisocial punishment, i.e., punishment directed at high contributors, has been observed especially in cross-cultural experiments (Herrmann, Thöni, and Gächter 2008). However, antisocial punishers seem relatively rare.

(Fischbacher, Gächter, and Fehr 2001).[3] Selfish individuals act according to the standard game theoretical prediction and always try to maximize their own wealth, without any concern for the payoff obtained by others. Conditional cooperators act according to the scheme of strong reciprocity. Assuming a population with heterogeneous preferences solves the second-order free-rider problem, because it can be expected that, in every group, there will be some conditional cooperators whose preference is to punish free-riders and to reward contributors. When sanctions are not allowed, selfish players succeed in driving cooperation down: as conditional cooperators face some defectors, their willingness to contribute declines and finally stops, since defection is the only available means to retaliate against defectors. By contrast, when sanctions are allowed, they are commonly used by conditional cooperators to force selfish actors to contribute (Fehr and Fischbacher 2003; Fehr and Gintis 2007). A symmetric argument can be made for rewards: they will typically be offered by conditional cooperators to high contributors. However, enticed by the possibility to gain rewards, selfish actors will then raise their contribution level, too. Therefore, on the macro level, we state the following hypothesis:

> *H1: Cooperation levels are higher when sanctions (carrots, sticks, or both) can be used than when they are not allowed (baseline).*

When carrots and sticks have to be compared, it is less clear what happens as contrasting arguments can be made favoring both sticks and carrots. Therefore, after summarizing the arguments we will present contrasting hypotheses for the effects of carrots versus sticks on cooperation. The effects of sticks are relevant especially for selfish actors, since they are normally used by conditional cooperators against free-riders. Conditional cooperators are unlikely to be punished, by definition. Conversely, selfish actors are likely to get punished for attempting to free-ride. As defectors get punished, the benefit earned by free-riding is eroded. Therefore, contributing becomes the payoff-maximizing choice. Sticks have thus the effect of increasing contribution of selfish individuals in future rounds of the PGG. On the other hand, carrots are likely to be given especially to conditional cooperators, as they should be the highest contributors. However, carrots affect both conditional cooperators and selfish actors. The former have two motives to contribute: first, they act according to strong reciprocity. Thus, if others cooperate they should cooperate as well. In addition, they may cooperate because they expect to receive rewards for high contributions. As for the selfish actors, carrots create economic advantages: if they are used extensively, they might produce higher payoffs (Rand et al. 2009). Thus, selfish individuals are only motivated to cooperate by the second

3 Fischbacher, Gächter, and Fehr (2001) conducted a study specifically designed to elicit individual preferences. They found that 50% of their participants were conditional cooperators and 30% were selfish types. The remaining 20% displayed other less easily interpretable behavioral patterns.

reason. That is, they will cooperate in order to receive monetary rewards and improve their payoff, if the probability of getting rewarded is high enough.

In addition to the economic argument, there is also a psychological argument favoring the efficacy of carrots over sticks. There is extensive evidence that, next to the positive effect of raising contribution, punitive sanctions produce some pernicious effects: they reduce mutual trust and trustworthiness within groups (Fehr and Rockenbach 2003; Mulder et al. 2006), they increase selfishness and hostility (Nikiforakis 2008), and they crowd out intrinsic motivation (Bohnet and Baytelman 2007). These arguments lead us to present the following hypothesis:

H2a: Cooperation levels are higher when only carrots can be used, compared to when only sticks can be used.

Next to the perspectives that predict advantages of carrots, there are also arguments for advantages of sticks. Dari Mattiacci and De Geest (2009) argue that punishment produces a "multiplication effect": while the sheer *opportunity* to punish is enough to sustain cooperation, punishment may never need to be *actually* used. In other words, assuming that sticks function as credible threats, if everybody cooperates – whether because they are cooperators, or because they are selfish but fear punishment – sticks never need to be applied. An effective stick can provide incentives for the same individual in different periods or for several individuals simultaneously, without anybody ever having to pay its costs. By contrast, carrots need to be used (and paid) each time a subject cooperates. Thus, carrots provide a more expensive means to sustain cooperation.

In addition to this economic argument, there are also psychological arguments favoring sticks over carrots. Individual's utility does not only depend on material benefits: people do not only avoid punishment and seek rewards for financial reasons, but also for the intrinsic benefit that receiving rewards and avoiding punishment provide. According to Lindenberg (2001) and Stigler and Becker (1977), individuals' universal goals include social approval, status, and affection. In particular, as they strive for social approval, people are sensitive to negative and positive sanctions. They strive to receive positive social reactions and avoid negative ones. However, social psychologists have shown that the desire to avoid negative feedbacks is typically stronger than the desire to receive positive feedbacks and this negative asymmetry, due to which "bad is stronger than good" is present across a broad range of phenomena (Baumeister et al. 2001). These arguments lead us to present the following hypothesis:

H2b: Cooperation levels are higher when only sticks can be used, compared to when only carrots can be used.

Since we have no clear prediction on the contribution level when both kinds of sanctions are allowed simultaneously, we turn to the results of Sefton, Schupp, and Walker (2007) and Rand et al. (2009) for comparison. These two studies compared costly rewards and costly punishment and yielded contrasting results, due to

crucial differences in the experimental design. In Sefton, Schupp, and Walker (2007) subjects played series of ten PGGs in groups of four. Each PGG was followed by a second stage in which they received extra points that they could use to punish, reward, and punish or reward (depending on the experimental treatment) the other group members. The ratio between cost and effect of the sanction was 1:1, i.e., one point spent on punishment decreased the income of the person punished of one point and one point spent on reward increased the income of the person rewarded of one point. Because of this feature, in Sefton, Schupp, and Walker (2007) rewards consisted in pure zero-sum horizontal transfer of points, without efficiency gain. In Rand et al.'s (2009) setup the game sequence was the same, but the total number of PGGs to be played was unknown to subjects. In addition, Rand et al. (2009) did not provide the subjects with extra money for the sanctioning part, but sanctions had to be paid from the money earned in the previous PGG. The cost to effect ratio of sanctions was 4:12, i.e., subjects could spend four points on punishment/reward to cause a decrease/increase of twelve points to the target. Finally, unlike in Sefton, Schupp, and Walker (2007), not only the group composition but also the subjects' IDs were held fixed throughout the experiment. Thus, mutual exchange of punishment or rewards – even unrelated to the behavior in the contribution stage of the game – was possible in Rand et al. (2009).

In Sefton, Schupp, and Walker (2007) the treatment in which both sticks and carrots were allowed yielded slightly higher contribution levels than any carrots and sticks separately, while the treatments with rewards only and punishment only produce similar levels of contribution to the public good. By contrast, in Rand et al. (2009) the contribution level was similar in the three treatments, but final payoffs were significantly higher in the two treatments where rewards were possible (i.e., one with only rewards and one with both rewards and punishments) than in the treatment where only punishment was possible. The high efficiency of the treatments with rewards is clearly due to the fact that, as personal IDs of the players were fixed in Rand et al. (2009), actors could exchange rewards mutually and these exchanges were highly productive due to the 4:12 cost to effect ratio. However, the possibility for mutual exchange of rewards effectively transforms the sanctioning stage of the game in a repeated prisoner's dilemma, for which mutual cooperation is a possible equilibrium, irrespectively of what happens in the PGG (Milinski and Rockenbach 2012). In addition, others have shown that, under certain conditions, the possibility of mutual exchanges of rewards could even be detrimental for the production of collective goods, because actors might increase each other's payoffs by exchanging gifts and stop contributing to the public good (Flache 2002).

As detailed in the experimental design section, we adopted Sefton, Schupp, and Walker's (2007) setup, with two modifications: first in order to strengthen the effect of sanctions we increased the magnitude of sticks and carrots by making the cost to effect ratio 1:2 instead of 1:1. Thus, in our study one point spent on sanctions increased or decreased the payoff of the target of two points. Second, we made the series of

rounds longer, in order to allow more time for the effects of sanctions to build up, but we turned the game into an indefinitely repeated game by making the end uncertain for the subjects. Note that we kept the group composition constant throughout the rounds but, unlike Rand et al. (2009), we did only inform subjects about the total number of points all others together used to sanction them, while we did not reveal the identities of these others. In this way, we make mutual exchanges of rewards (or mutual punishment) impossible. Given that our setup largely resembles the one adopted by Sefton, Schupp, and Walker (2007), except for increasing the potential efficacy of sanctions by modifying the cost to effect ratio, we expect to replicate the effect found in their study. Therefore, we propose the following hypothesis:

H3: If carrots and sticks are allowed simultaneously, the cooperation levels are higher than when only either of the two is allowed.

18.2.4 Hypotheses on group solidarity

While previous research has focused primarily on the necessity of sanctioning institutions to promote cooperation in public goods, possible negative side-effects of these institutions have received considerably less attention. However, Tenbrunsel and Messick (1999) argued that the use of monetary sanctions can have perverse effects on group solidarity because it imposes an economic frame, which may crowd out intrinsic motivation and lead subjects to cooperate only to avoid financial losses. The consequences of using carrots and sticks in terms of solidarity among group members have not yet been sufficiently investigated. However, the existence of important side-effects, potentially affecting group solidarity, has been noted by several researchers (Oliver 1980; Fehr and Rockenbach 2003; Mulder et al. 2006, Irwin, Mulder, and Simpson 2014). The existing literature suggests that carrots can promote solidarity and emotional attachment to the group (Markovsky and Lawler 1994; Friedkin 2004). Sticks, on the other hand, may succeed in controlling defection of group members, but could have detrimental side effects on solidarity due to the negative emotional responses they evoke (Oliver 1980; Friedkin 2004). These side effects can range from unhappiness to tension and hostility within the group (Oliver 1980). In general, punishers are disliked while rewarders are liked (Friedkin 2004). Moreover, sticks may crowd out the intrinsic motivation to help the other group members, and damage or undermine mutual trust (Fehr and Rockenbach 2003; Mulder et al. 2006). More specifically, being punished by another group member decreases trust and willingness to cooperate with the punisher (Oliver 1980; Fehr and Rockenbach 2003; Mulder et al. 2006).

As stated above, gene-culture coevolution models of altruistic preferences attribute an important role to cultural group selection in the process leading to the evolution of complex sanctioning institutions. Therefore, intrinsic motivation to support

the group as well as mutual trust and solidarity between group members are very important factors in the competition between groups. Low levels of mutual trust and solidarity might lead groups to perform poorly in competition with other groups. Demotivated group members might be tempted to leave the group. Ultimately, in the long-term a group characterized by an unpopular sanctioning institution might be defeated by other groups or simply dissolve.

The same argument holds for modern groups whose members stay the same for a long time, as it is the case for many small groups (e.g. clubs and organizations). If the members do not feel attached to the group or dislike each other, they will ultimately be less willing to contribute, because feelings have the ability to weaken or strengthen the bonds between group members (Markovsky and Lawler 1994). Although investigating long-term effects in the artificial setting of a computer-mediated laboratory experiment is certainly difficult, our experiment aims at comparing whether carrots or sticks lead to more cooperation in longer sequences of games than normally studied. In addition, we investigate whether carrots lead to more group solidarity by letting subjects play an additional one-shot "person-to-group" Dictator Game (DG), after the repeated PGGs. This one-shot DG – explained in details in the experimental design section – constitutes our individual measurement of group solidarity. Therefore, assuming that actors' attachment and solidarity to the group is affected by the sanctions that they receive, we postulate the following hypotheses:

H4: Individuals show lower solidarity towards other group members, the more they have been punished.

H5: Individual show higher solidarity towards other group members, the more they have been rewarded.

18.3 The experimental design

To test the hypotheses a series of computerized experiments was conducted with z-Tree (Fischbacher 2007) in the Experimental Laboratory for Sociology and Economics of a large Dutch University. Eight sessions took place at the end of 2009 with a total of 152 subjects. The subjects were students from various faculties, recruited via the online recruiting system ORSEE (Greiner 2004).

Like in Sefton, Schupp, and Walker (2007), the PGG was played as a repeated game. This characteristic allows for reciprocity in the sense that subjects can react to their fellow group members' investments either directly by sanctioning (punishing or rewarding, depending on the treatment), or indirectly, by adjusting their own investments accordingly in the next round. Furthermore, to simulate cooperation problems in groups as realistically as possible, the subjects did not know in advance when the game would end, since the end of a repeated interaction is generally unknown in real

life situations. The interactions were anonymous and, unlike in Rand et al. (2009), upon receiving a sanction subjects could not identify who had sanctioned them, so that mutual rewarding (or punishing) between rounds was impossible.

Each of the eight experimental sessions lasted between 45 and 75 minutes, depending on the treatment. Each session consisted of two series of repeated PGGs (thus, two super-games), played in the same experimental treatment. At the beginning of each session, subjects played two trial rounds to get acquainted with the procedures of the game. In these trial rounds, which had no influence on the subjects' earnings, the behaviors of the others in each group were pre-programmed. In the first trial round, the subject was informed that the other (virtual) group members had invested four points each, in the second trial round six points each. The pre-programming took place to guarantee that all subjects started the experimental rounds with the same experience. The lowest and highest possible investments, such as zero or ten points, were avoided to prevent any suggestions for extreme actions. When all individuals had completed the trial rounds, they were randomly matched into groups of four and the second part of the experiment began. The first of two super-games was played, which consisted of between 20 and 30 rounds of PGGs. After round 21 the game stopped with a probability of 10%, after round 22 with a probability of 20%, and so on until the game would stop with certainty in round 30.

Each round consisted of two parts, a PGG stage, followed by a sanctioning stage, except for the control treatment that had no sanctioning stage. At the beginning of the PGG stage, the subjects received an initial endowment of ten points and they could decide how much of this endowment to invest into a group account. Whatever the subjects did not invest into the group account was stored directly into the subject's individual account to be paid eventually at the end of the experiment. The amount collected in the group account was multiplied by three and then equally distributed between the four group members. In the sanctioning stage, the subjects received five additional points, which they could use, at a cost to effect ratio of 1:2, to punish and/or reward the other group members. That is, if a subject spent, for example, one point of his/her five points on punishing/rewarding someone else, two points were subtracted from/added to the target's payoff. All points that were not spent in sanctions were added to the subject's individual account. This manipulation of the sanctioning option is similar to the one adopted by Sefton, Schupp, and Walker (2007). However, we made sanctions more cost-effective, by adding/subtracting two points, instead of one, per every point spent on reward/punishment. This modification was implemented in order to better test our hypothesis concerning the advantage of carrots over sticks (Hypothesis 2a), because the main argument supporting that hypothesis is that rewards can increment earnings. However, as stated above, in Sefton, Schupp, and Walker (2007) rewards were simply a zero-sum transfer of points between players and, therefore, they were not economically attractive.

The experiment had four treatments and we conducted two sessions per treatment. The first treatment was the Baseline Treatment, which had no options to sanction.

The second was the Sticks Treatment, in which, in the sanctioning stage, subjects could spend any of their five points available for sanctions on punishments only. In the third treatment (Carrots Treatment) all group members could spend any of their five points available for sanctions on rewards only. In the fourth treatment both punishment and rewards were possible (Sticks & Carrots Treatment). However, the endowment at disposal remained at five points in total for both kinds of sanctions. Each subject could not spend more than the five points provided for punishing and/or rewarding. The individuals could not use their own earnings to sanction others because otherwise, subjects who earned more in the PGG might have been more willing to spend points on punishing and/or rewarding others than subjects who earned little. Thus, the measure of how willing people are to punish defectors and reward cooperators would have been biased by the amount earned in the PGG. After finishing the first super-game, all subjects were randomly re-matched in different four-person groups and the second super-game started. Throughout each super-game, the subjects were informed about the individual investments of the other members of their group by a table located at the lower part of the computer screen. Thus, subjects could make their sanctioning decisions contingent on the contributions of their fellows groups members. Furthermore, the subjects were informed about the amount that they had invested to inflict sanctions on others in each round, and about the losses and gains that were inflicted on them by punishments and rewards received from the other group members (if applicable).

Twenty subjects took part in each session, except for two sessions which had only 16 subjects (one in the Baseline and one the Carrots Treatment). Therefore, five (or four) groups of four persons were formed in each session. Each subject was seated in front of a computer station, which was visually separated from the others. As a start, the instructions about the procedures of the experiment were handed out, the subjects could choose between English and Dutch.[4] Of the 152 subjects 104 (68.4%) chose the Dutch instructions and 48 (31.6%) the English version.

Each of the two super-games per session was directly followed by a person-to-group DG which serves as our behavioral measure of a subject's solidarity towards the group. The person-to-group DG works as follows: each subject receives an endowment of 40 points. Then, he or she can decide to give some of these 40 points to the group. The amount chosen is equally distributed between all three remaining group members, without being multiplied by any number beforehand. The subjects may choose to give zero points to their fellow group members, in which case the 40 points are added to their own earnings. We used the amount given as an individual measure of solidarity towards one's own group. Finally, the experiment ended with a questionnaire asking for demographic characteristics (such as gender, age, nationality, field of studies, amount of money at disposal each month).

4 Instructions can be obtained contacting the authors.

The earnings of each subject were determined by the contributions accumulated in the group account – in every round of the PGG – divided by four (all group members) plus the rest of the endowments (from the PGG and the sanctioning part of each round) that subjects did not invest in the PGG or in sanctions, plus the points received as rewards and minus the points lost due to punishment. In addition, the points kept as well as the points received by fellow group members in the DG were likewise added to the individual account. At the end of the experiment, the subjects were paid in Euros at an exchange rate of 125 points = 1 Euro.

Of the 152 subjects who participated, 36 played in the Baseline Treatment, 40 in the Sticks Treatment, 36 in the Carrots Treatment, and 40 in the Sticks & Carrots Treatment. In total, 7568 decisions were recorded; 304 of these in the trial rounds. Thus, we recorded 7264 decisions that had influence on the subjects' earnings. Between 21 and 27, on average 23.9 rounds, were played in all super-games. The subjects earned between 5.5 and 16.5 Euros, the lowest earnings were received in the Baseline Treatment. The average profit was 12.60 Euros.[5]

18.4 Methods

For the hypotheses on cooperation, the average of the *contributions* per group serves as the dependent variable. The four different treatments serve as independent variables. For the analysis, we estimated a panel regression to account for the repeated observations within subject over time. We only used the first 20 rounds of the super-games to discard possible endgame effects. We did alternative analyses such as (1) including the final rounds, (2) using an interval regression model to take into account that groups could not contribute less than 0 or more than 40, and (3) only analyzing the first super-game. In all these analyses the pattern of results shown below is robust.

The dependent variable *solidarity* is measured by the individual contributions to the group in the *DGs*, following every super-game. The four different treatments serve as independent variables. The other independent variables are: the extent to which individuals contributed to the public good on average, the individual average received punishments, and individual average received rewards. To test the hypotheses on group solidarity we estimated a non-hierarchical mixed effects model, which includes random effects for groups and subjects. This mixed effects model is cross-classified because most subjects are involved in two different groups. The model allows us to control for possible effects of individual differences and group differences.

5 Due to a programming error, rematching was unsuccessful in two sessions. Therefore, the second super-game in these two sessions had to be excluded from the analyses presented below.

18.5 Results

18.5.1 Descriptive results

Over all treaments the contribution level was on average quite high: 8.2 points out of ten possible. Looking at the contribution levels in the different treaments, it is clear that cooperation was higher in the treatments in which sanctions were possible compared to the Baseline Treatment. The Carrots Treatment had the lowest average contribution level among the treatments with sanctions. In the two treatments with punishment, subjects assigned on average 0.35 and 0.38 punishment points to the three other subjects. Actually, in 87% of the rounds in which they could punish, subjects did not punish at all. Rewards were more frequently given, namely in 66% of the rounds in which subjects could give rewards. The average amount of points that subjects used for sanctions was 2.37 points in the Carrots Treatment and 2.03 points in the Sticks & Carrots Treatment. In the DG, the amount given to the other group members was on average 11.6 of 40 possible points, which is relatively low, compared to the percentages invested in the PGG. Table 18.1 summarizes the descriptive statistics.

Table 18.1: Descriptive statistics.

	Range	Mean	S.D.	Number of obs.
Number of rounds played	12–27	24.11	1.75	68
Contribution level in groups (overall)	0–40	36.74	6.43	1640
Individual contributions (overall)	0–10	9.18	2.16	6560
Contribution in Baseline	0–10	8.15	3.23	1384
Contribution in Sticks	0–10	9.69	1.26	1920
Contribution in Carrots	0–10	8.80	2.36	1336
Contribution in Sticks & Carrots	0–10	9.69	1.23	1920
Punishment given in Sticks	0–5	0.38	1.18	1920
Punishment given in Sticks & Carrots	0–5	0.35	0.97	1920
Punishment received in Sticks	0–12	0.38	1.09	1920
Punishment received in Sticks & Carrots	0–14	0.35	1.04	1920
Rewards given in Carrots	0–5	2.37	1.82	1336
Rewards given in Sticks & Carrots	0–5	2.03	1.89	1920
Rewards received in Carrots	0–8	2.37	1.56	1336
Rewards received in Sticks & Carrots	0–9	2.03	1.62	1920
Contributions in Dictator Game	0–40	11.63	13.74	268
Age	17–40	21.44	3.00	152
Earnings	5.5–16.5	12.60	2.72	152

As an overview, the contribution levels in each treatment are shown in Figure 18.1. This figure shows the average contribution level in each round of the 68 groups of four subjects who participated in the study. Figure 18.1 makes evident that contributions are the highest in the treatments that include sticks. Overall, the contributions in this experiment are considerably larger than in other public good experiments. This might be due to the large number of rounds subjects played. As mentioned in the description of the experiment, the super-games ended at random between 20 and 30 rounds. Therefore, the dots corresponding to rounds after the 20th represent fewer observations. For descriptive purposes we still show these rounds. They show that there are no endgame effects at all for the treatments that include sticks, while the contributions seem to decline after round 20 in the other two treatments. Furthermore, Figure 18.1 shows that the subjects in all treatments started with relatively high contributions compared to earlier experiments and even rose thereafter. However, only in the Sticks and Sticks & Carrots Treatments very high contribution levels, up to full cooperation, were maintained until the end. In the Carrots Treatment the cooperation levels were lower, and in the Baseline Treatment the contributions fell, after an initial rise, even below the starting level of cooperation.

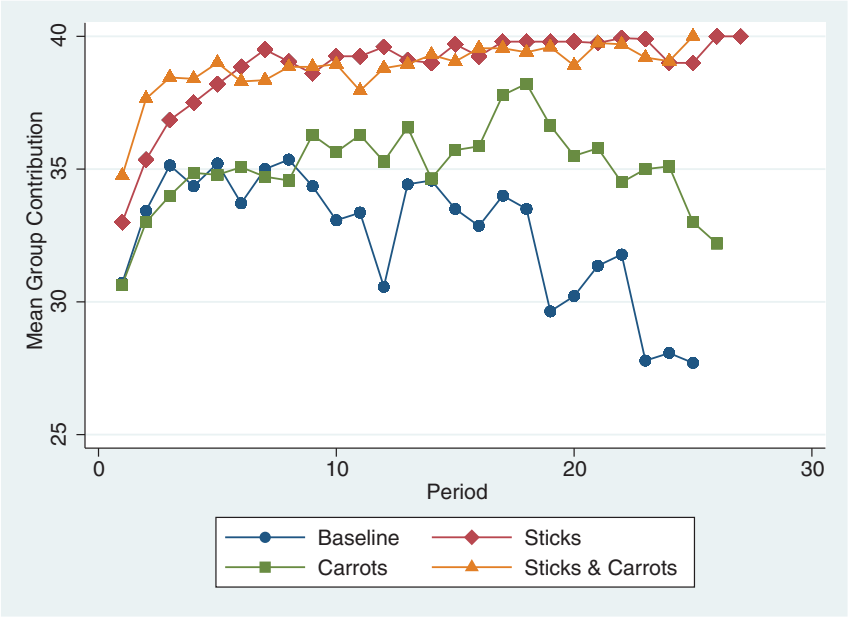


Figure 18.1: Contribution levels in the groups by treatment.

If we analyze the two super-games separately, we find that in the second super-game the subjects start in all treatments with higher contributions than in the first super-game.

Moreover, in the second super-game almost perfect cooperation is reached quicker in the Sticks and Sticks & Carrots Treatments and is stable (from round 5 onwards). The contribution levels in the Carrots and the Baseline Treatments are likewise higher than in the first super-game, but still lower than in the treatments where punishment is possible. Summarizing, the few changes that we implemented compared to the setup used by Sefton, Schupp, and Walker (2007), namely increasing the cost to effect ratio of sanctions from 1:1 to 1:2, letting subjects play more rounds, and making the end of the game probabilistic, apparently made the reputation argument in repeated games more salient, driving up contribution levels.

18.5.2 Explanatory results

18.5.2.1 Contribution to the public good

The panel regression on cooperation levels shows that contributions were significantly higher in the two treatments in which punishment was allowed (*sticks* and *sticks & carrots*) compared to the *baseline* where no sanction was allowed (see Table 18.2, Model 1). This effect is not significant for the Carrots Treatment, although the coefficient is

Table 18.2: Two-level regression on contribution with clustering on groups.

	Model 1		Model 2	
	Coefficient	**s.e.**	**Coefficient†**	**s.e.**
Baseline (reference)				
Sticks	5.21**	1.54	5.04**	1.54
Carrots	1.95	1.67	1.79	1.67
Sticks & Carrots	5.28**	1.54	5.16**	1.63
Round (Baseline)			–0.11**	0.04
Round (Sticks)			0.22**	0.03
Round (Carrots)			0.21**	0.04
Round (Carrots & Sticks)			0.12**	0.03
Constant	33.35**	1.18	34.53**	1.25
Random part				
Group level (st. dev.)	4.34		4.35	
Obs. level (st. dev.)	3.83		3.69	
R^2 within groups	0.00		0.07	
R^2 between groups	0.21		0.21	
R^2 overall	0.13		0.16	
N (groups)	*68*		*68*	
N (observations)	*1360*		*1360*	

Notes: * $p<0.05$, ** $p<0.01$ (two-sided)
† Main effects of conditions calculated for round 10

positive. Thus, Hypothesis 1 is only partly supported: cooperation levels are higher if negative sanctions can be used, but not if only positive sanctions can be used. If we include the time trends for each treatment (Model 2), we see that the differences between treatments are mainly due to increased contributions in the treatments with sanction opportunities and most strongly in treatments with sticks, while contributions decreased in the Baseline Treatment. Including other control variables, such as age and gender, does not change much in the model, but it only slightly reduces the effect sizes of the dummies for *sticks* and *sticks & carrots* and of the round variables. Thus, the model including the controls is not shown here.

Concerning Hypotheses 2a and 2b, about the comparison of contribution levels between *sticks* and *carrots*, the former is rejected and the latter is supported. That is, contribution levels are significantly higher when only punishments can be used, compared to when only rewards can be used. Finally, there are no significant differences in the contribution level between the Sticks Treatment and the Sticks & Carrots Treatment. However, the difference between the Carrots Treatment and the Sticks & Carrots Treatment is significant ($p = 0.0310$). Thus, Hypothesis 3 is only partly confirmed: if punishments and rewards are allowed simultaneously the contribution level is significantly higher than when only rewards are allowed, but it is not different from the contribution achieved when only punishment is available. Thus, the final message is clearly that sticks are more effective than carrots at sustaining cooperation and when sticks are available, the contribution level does not change, whether carrots are also available as well or not.

18.5.2.2 Solidarity

The amount given in the person-to-group DGs is used to operationalize solidarity. In Table 18.3, Model 1 shows no significant differences between any of the treatments. In Model 2, though, we can see that subjects gave significantly less to the group in the DG if they played the previous super-game in the Carrots Treatment *and* they did not receive any rewards. However, if individuals did receive rewards, they significantly gave more to the other group members in the DG. These two effects cancel out each other, because there is no main effect of carrots on solidarity. Apparently, subjects expect to receive rewards if rewards are possible. If they do not obtain carrots when they could, their solidarity towards the group diminishes. To compensate for this, subjects should receive 2.12 (= −7.47/3.52) points of rewarding per round. From Table 18.1 we know that subjects indeed on average gave a little more than 2 units in rewards per round. This explains why the main effect is indeed about zero. We did some additional analyses to see whether the experiences in the latest rounds might have had a relatively large effect on the behavior in the DG. Therefore, we distinguished between the experiences in the first 20 rounds with

Table 18.3: Mixed effects regression on contribution in Dictator Game.

	Model 1		Model 2	
	Coefficient	**s.e.**	**Coefficient**	**s.e.**
Baseline (reference)				
Sticks	3.08	2.98	1.70	3.09
Carrots	1.53	3.13	−7.47*	3.55
Sticks & Carrots	2.67	2.98	−5.91	3.43
Average contribution			0.18	0.16
Average received punishments			0.47	1.45
Average received rewards			3.52**	0.81
Constant	9.51**	2.21	3.63	5.46
Random effects				
Group level (st. dev.)	1.70		0.00	
Subject level (st. dev.)	10.94		10.53	
Obs. Level	8.03		7.92	
Log restricted-likelihood	−1041.52		−1029.43	
N (groups)	*68*		*68*	
N (individuals)	*152*		*152*	
N (observations)	*268*		*268*	

Note: *p<0.05, **p<0.01 (two-sided)

the experiences in the last couple of rounds. There were no significant differences between these effects.

Therefore, we conclude that Hypothesis 4 cannot be confirmed, since the results on the Sticks Treatments are not significant. Punishments do not lead to a negative attitude towards the group and there is not less solidarity in the punishments treatments. One explanation for this is that punishments were hardly used and, in line with the "multiplication effect" argument (Dari Mattiacci and De Geest 2009), cooperation was very high in both treatments with punishment. Thus, punishment worked as a credible threat. Conversely, Hypothesis 5 is partly supported: the individual solidarity towards the group is heightened by actual received rewards. It appears that the sheer possibility of receiving rewards leads to the expectation that one *should* receive rewards. If this expectation is not fulfilled, cooperation even decreases in the treatment with rewards.

18.6 Discussion and conclusion

The aim of this chapter was to investigate whether positive or negative sanctions in public good type of situations are more effective at promoting cooperation. Moreover,

We looked at the consequences that positive and negative sanctions produce in terms of group solidarity. We addressed these two research problems by conducting a computerized experiment. Our main findings show that cooperation levels are generally higher if sticks or sticks and carrots are allowed (Hypothesis 1). We also found that average contribution levels were rather high in our study. Moreover, contribution levels were higher in the condition in which only sticks were allowed than in the condition in which only carrots were allowed (Hypothesis 2b). When both sanctions were allowed simultaneously (Hypothesis 3), the contribution levels were significantly higher than in the condition with rewards only, but similar to the condition with punishments only.

The results on group solidarity do not lend support to the hypothesis that individuals care less about their group after they are punished (Hypothesis 4). However, if the subjects receive rewards they indeed display a higher group solidarity, as they give more to the group in the DG (Hypothesis 5). By contrast, the mere possibility to use sanctions, whether positive or negative, does not have an effect on group solidarity. Surprisingly, receiving punishment does not affect group solidarity. However, this result might be due to the low amount of actual punishment observed in our experiment. Thus, our results do not allow a strong test of the potentially demotivating effects of implemented punishments. Conversely, rewards were used more often and, as expected, produced the effect of increasing group solidarity. In addition, subjects who did not receive rewards in the condition where rewards were possible exhibited lower solidarity to the group than subjects in other conditions. Consequently, our data do not support the argument that carrots can sustain cooperation in social dilemmas (Rand et al. 2009), nor the argument that sticks have detrimental side effects in terms of group solidarity (cf. Fehr and Rockenbach 2003; Mulder et al. 2006). However, the finding that receiving rewards increases group solidarity gives some indication that cultural norms based on rewards may benefit the group in the long term, if we assume that higher group solidarity makes groups more resilient in the context of intergroup competition (cf. Richerson, Boyd, and Henrich 2003).

Our experimental setup was designed to simultaneously address some limitations of the designs used by Rand et al. (2009) and Sefton, Schupp, and Walker (2007). Our results are consistent with the idea that the high payoffs in a PGG with rewards found in Rand et al. (2009) could be due to the mutual exchange of highly efficient rewards, rather than being a pure effect of rewards on cooperation (Milinski and Rockenback 2012). By contrast, as we made the (monetary) consequences of sanctions bigger than in Sefton, Schupp, and Walker (2007), rewards were used more often than punishment by our subjects while punishment was used more often in Sefton, Schupp, and Walker (2007). Yet, we found higher contribution levels in the punishments only treatment than in the rewards only treatment, suggesting that sticks function as credible threats, i.e., they support cooperation without needing to be used (cf. Dari Mattiacci and De Geest 2009). Our experimental design also implemented a higher number of within group interactions than did Sefton, Schupp, and

Walker (2007), coupled with an uncertain end of the super-game. As a result, we obtained relatively high levels of cooperation, even in the treatment without sanctions. Consequently, our subjects provided more room for the use of positive than negative reinforcers. Due to the 1:2 cost to effect ratio, in our rewards condition, Pareto-superior outcomes could be reached if everyone gave rewards to everyone all the time. Perhaps some of our subjects realized this and others did not. If the actors who did realize this expected that indeed everyone exchanged rewards, they may have been disappointed when others did not conform to this expectation, because they lost points from providing rewards that they did not receive back from others. This last point is clearly related to the costs of rewards. The actors who provided rewards would not have been so much bothered if the rewards were for free. This implies that the cost argument of Dari-Mattiacci and De Geest (2009) is relevant in this context. The threat of sticks is effective and cheap, while the costs of continuously providing carrots make them less effective.

Our results also imply that more theoretical work about group solidarity is needed. For example, the negative effect of sticks, e.g., reduced group solidarity, might not be generated by the possibilities of punishments alone, but it might require a context in which punishment is necessarily applied. This could be modeled as a public good with uncertainty, i.e., a setting in which the contribution to the public good is not always perfectly visible and sometimes actors do not seem to contribute, while they actually do. Finally, the weak results concerning the effects of sanctions on solidarity might also be due to our measurement of group solidarity as a one-shot person-to-group DG. The behavior in the DG can be determined by many subtle cues in the PGGs. In future studies, it would be advisable to complement the findings with more detailed measurements of solidarity and attachments to the group, for example, using some attitudinal scales, next to behavioral measures such as the DG.

References

Andreoni, James. 1988. "Why Free-Ride?: Strategies and Learning in Public Goods Experiments." *Journal of Public Economics* 37: 291–304.

Andreoni, James, William Harbaugh, and Lise Vesterlund. 2003. "The Carrot or the Stick: Rewards, Punishments, and Cooperation." *American Economic Review* 93: 893–902.

Axelrod, Robert. 1984. "*The Evolution of Cooperation*." New York: Basic Books.

Balliet, Daniel, Laetitia Mulder, and Paul A. M. van Lange. 2011. "Reward, Punishment, and Cooperation: A Meta-Analysis." *Psychological Bulletin* 137: 594–615.

Barclay, Pat. 2006. "Reputational Benefits for Altruistic Punishment." *Evolution and Human Behaviour* 27: 325–344.

Boyd, Robert J., Herbert Gintis, Samuel Bowles, and Peter J. Richerson. 2003. "The Evolution of Altruistic Punishment." *Proceedings of the National Academy of Science* 100: 3531–3535

Baumeister, Roy F., Ellen Bratslavsky, Catrin Finkenauer, and Kathleen D. Vohs. 2001. "Bad Is Stronger Than Good." *Review of General Psychology* 5: 323–370.

Bohnet, Iris, and Yael Baytelman. 2007. "Institutions and Trust: Implications for Preferences, Beliefs and Behavior." *Rationality and Society* 19: 99–135.

Buskens, Vincent, and Werner Raub. 2013. "Rational Choice Research on Social Dilemmas: Embeddedness Effects on Trust." In *Handbook of Rational Choice Social Research*, edited by Rafael Wittek, Tom A. B. Snijders, and Victor Nee, 113–150. Stanford: Stanford University Press.

Calvert, Randall L. 1995. "Rational Actors, Equilibrium, and Social Institutions." In *Explaining Social Institutions*, edited by Jack Knight and Itai Sened, 57–94. Ann Arbor, MI: University of Michigan Press.

Chauduri, Ananish. 2011. "Sustaining Cooperation in Laboratory Public Goods Experiments: A Selective Survey of the Literature." *Experimental Economics* 14: 47–83.

Dari-Mattiacci, Giuseppe, and Gerrit De Geest. 2009. "Carrots, Sticks, and the Multiplication Effect." *Journal of Law, Economics and Organization* 73: 377–386.

Fehr, Ernst, and Urs Fischbacher. 2003. "The Nature of Human Altruism." *Nature* 425: 785–791.

Fehr, Ernst, and Simon Gächter. 2002. "Altruistic Punishment in Humans." *Nature* 415: 137–140.

Fehr, Ernst, and Herbert Gintis. 2007. Human Motivation and Social Cooperation: Analytical and Experimental Foundations. *Annual Review of Sociology* 33: 43–64.

Fehr, Ernst and Joseph Henrich. 2003. "Is strong Reciprocity a Maladaptation? On the Evolutionary Foundations of Human Altruism." In *Genetic and Cultural Evolution of Cooperation*, edited by Peter Hammerstein, 55–82. Cambridge, MA: The MIT Press.

Fehr, Ernst, and Bettina Rockenbach. 2003. "Detrimental Effects of Sanctions on Human Altruism." *Nature* 422: 137–140.

Fehr, Ernst, and Klaus, M. Schmidt. 1999. "A Theory of Fairness, Competition, and Cooperation." *Quarterly Journal of Economic* 114: 817–68.

Fischbacher, Urs. 2007. "z-Tree – Zurich Toolbox for Readymade Economic Experiments – Experimenter's Manual." *Experimental Economics* 10: 171–178.

Fischbacher, Urs, Simon Gächter, and Ernst Fehr. 2001. "Are People Conditionally Cooperative? Evidence from a Public Goods Experiment. *Economics Letters* 71: 397–404.

Flache, Andreas. 2002. "The Rational Weakness of Strong Ties. Failure of Group Solidarity in a Highly Cohesive Group of Rational Agents." *Journal of Mathematical Sociology* 26: 189–216.

Friedkin, Noah E. 2004. "Social Cohesion." *Annual Review of Sociology* 30: 409–25.

Gächter, Simon. 2013. "Rationality, Social Preferences and Strategic Decision-Making from a Behavioral Economics Perspective." In *Handbook of Rational Choice Social Research*, edited by Rafael Wittek, Tom A. B. Snijders, and Victor Nee, 33–71. Stanford: Stanford University Press.

Gintis, Herbert. 2000. "Strong Reciprocity and Human Sociality." *Journal of Theoretical Biology* 206: 169–179.

Greiner, Ben. 2004. "The Online Recruitment System ORSEE 2.0. A Guide for the Organization of Experiments in Economics." University of Cologne, Working Paper Series in Economics 10.

Hamilton, William D. 1963. "The Evolution of Altruistic Behavior." *The American Naturalist* 97: 354–356.

Herrmann, Benedikt, Christian Thöni, and Simon Gächter. 2008. "Antisocial Punishment across Societies." *Science* 319: 1362.

Hobbes, Thomas. 1651. *Leviathan or the Matter, Forme and Power of a Commonwealth, Ecclesiasticall and Civill*. Oxford: Clarendon Press, 1909, reprinted from the edition of 1651, London.

Irwin, Kyle, Laetitia Mulder, and Brent Simpson. 2014. "The Detrimental Effects of Sanctions on Intragroup Trust: Comparing Punishments and Rewards." *Social Psychology Quarterly* 77: 253–272.

Johnson, Dominic P., Pavel Stopka, and Stephen Knights. 2003. "The Puzzle of Human Cooperation." *Nature* 421: 911–912

Keser, Claudia, and Frans van Winden. 2000. "Conditional Cooperation and Voluntary Contributions to Public Goods." *Scandinavian Journal of Economics* 102: 23–39.

Kreps, David M., Paul Milgrom, John Roberts, and Robert Wilson. 1982. "Rational Cooperation in the Finitely Repeated Prisoners' Dilemma." *Journal of Economic Theory* 27: 245–252.

Ledyard, John O. 1995. "Public Goods: A Survey of Experimental Research." In *Handbook of Experimental Economics*, edited by John H. Kagel and Alvin E. Roth, 111–194. New York: Princeton University Press.

Lindenberg, Siegwart. 2001. "Social Rationality versus Rational Egoism." In *Handbook of Sociological Theory*, edited by John H. Turner, 635–668. New York: Kluwer Academic/Plenum Publishers.

Markovsky, Barry, and Edward J. Lawler. 1994. "A New Theory of Group Solidarity." *Advances in Group Processes* 11: 113–137.

Masclet, David, Charles Noussair, Steven Tucker, and Marie-Claire Villeval. 2003. "Monetary and Nonmonetary Punishment in the Voluntary Contributions Mechanism." *American Economic Review* 93: 366–380.

Milinski, Manfred, and Bettina Rockenbach. 2012. "On the Interaction of the Stick and the Carrot in Social Dilemmas." *Journal of Theoretical Biology* 299:139–143.

Mulder, Laetitia B., Eric Van Dijk, David De Cremer, and Henk A. M. Wilke. 2006. "Undermining Trust and Cooperation: The Paradox of Sanctioning Systems in Social Dilemmas. *Journal of Experimental Social Psychology* 42: 147–162.

Nikiforakis, Nikos. 2008. "Punishment and Counter-punishment in Public Goods Games: Can We Really Govern Ourselves?" *Journal of Public Economics* 92: 91–112.

Nowak, Martin A. 2006. "Five Rules for the Evolution of Cooperation." *Science* 314: 1560–1563.

Nowak, Martin A, Anna Dreber, David G. Rand, and Drew Fudenberg. 2008. "Winners Don't Punish." *Nature* 452: 348–351.

Nowak, Martin A., and Karl Sigmund. 1998. "Evolution of Indirect Reciprocity by Image Scoring." *Nature* 393: 573–577.

Oliver, Pamela. 1980. "Rewards and Punishments as Selective Incentives for Collective Action: Theoretical Investigations." *American Journal of Sociology* 58: 1356–1375.

Olson, Mancur. 1965. *The Logic of Collective Action: Public Goods and the Theory of Groups*. Cambridge: Harvard University Press.

Pan, Xiaofei, and Daniel Houser. 2017. "Social Approval Competition and Cooperation." *Experimental Economics* 20: 309–332.

Rabin, Matthew. 1993. "Incorporating Fairness into Game Theory and Economics." *American Economic Review* 83: 1281–1302.

Rand, David G., Anna Dreber, Tore Ellingsen, Drew Fudenberg, and Martin A. Nowak. 2009. "Positive Interactions Promote Public Cooperation." *Science* 325: 1272–1275.

Raub, Werner, Vincent Buskens, and Rense Corten. 2015. "Social Dilemmas and Cooperation." In *Handbuch Modellbildung und Simulation in den Sozialwissenschaften*, edited by Norman Braun and Nicole J. Saam, 597–626. Wiesbaden: Springer VS.

Raub, Werner, and Jeroen Weesie. 1990. "Reputation and Efficiency in Social Interactions: An Example of Network Effects. *American Journal of Sociology*, 96: 626–654.

Richerson, Peter J., Robert T. Boyd, and Joseph Henrich. 2003. "Cultural Evolution of Human Cooperation." In *Genetic and Cultural Evolution of Cooperation*, edited by Peter Hammerstein, 357–388. Cambridge, MA: MIT Press.

Sefton Martin, Robert Shupp, and James M. Walker. 2007. "The Effects of Rewards and Sanctions in Provision of Public Goods." *Economic Inquiry* 45: 671–690.

Spence, Michael. 1973. "Job Market Signaling." *Quarterly Journal of Economics* 87: 355–374.

Stigler, George J., and Gary S. Becker. 1977. "De Gustibus Non Est Disputandum." *American Economic Review* 67: 76–90.

Stoop, Jan, Daan Van Soest, and Jana Vyrastekova. 2018. "Rewards and Cooperation in Social Dilemma Games." *Journal of Environmental Economics and Management* 88: 300–310.

Takacs, Karoly, and Bela Janky. 2007. "Smiling Contributions: Social Control in a Public Goods Game with Network Decline." *Physica A* 378: 76–82.

Tenbrunsel, Ann E., and David M. Messick. 1999. "Sanctioning Systems, Decision Frames, and Cooperation." *Administrative Science Quarterly* 44: 684–707.

Traulsen, Arne, and Martin A. Nowak. 2006. "Evolution of Cooperation by Multilevel Selection." *Proceedings of the National Academy of Science* 103: 10952–10955.

Trivers, Robert L. 1971. "The Evolution of Reciprocal Altruism." *Quarterly Review of Biology* 46: 35–57.

Willer, Robb. 2009. "Groups Reward Individual Sacrifice: The Status Solution to the Collective Action Problem." *American Sociological Review* 74: 23–43.

Yamagishi, Toshio. 1986. "The Provision of a Sanctioning System as a Public Good." *Journal of Personality and Social Psychology* 51: 110–116.

Zahavi, Amos. 1995. "Altruism as a Handicap – the Limitations of Kin Selection and Reciprocity." *Journal of Avian Biology* 26: 1–3.

Part III: **Field Studies**

Gerrit Rooks, Chris Snijders and Frits Tazelaar

19 A Sociological View on Hierarchical Failure: The Effect of Organizational Rules on Exchange Performance in Buyer-Supplier Transactions

Abstract: The classic Transaction Cost Economics view is that a key reason for firms to exist is that they offer a way to overcome problematic market transactions. If it is too complicated, expensive, or risky to buy a good on the market, consider hiring employees to make it in-house – especially if it is a good that you (and others) might need often. The implicit argument is that, for this reason, firms are a potentially rational response to less advantageous markets. However, firms are rational responses only when they themselves are organized in a way that is efficient enough to outperform the market. We consider firms' hierarchical efficiency by analyzing the existence and consequences of rules and procedures, effectively testing two competing arguments. On the one hand, rules and procedures are one way in which firms can achieve efficiency, through specialization and formalization of what a firm has learned. On the other hand, rules can be imprecise and rigid, a nuisance to deal with, and just coincidental traces of what has gone wrong in the past.

Using a database of more than 800 transactions in which German small and medium sized businesses buy ICT products and services, we consider the role of rules and procedures in a large-scale quantitative way. It turns out that rules show a pyramid-like structure where some firms have less and others have more codified rules. Our results furthermore suggest that rules might not be the clotted efficiency they have been argued to be. A high rule-density goes with increased investments in contracting ("thicker contracts") and not with decreased ex post transaction problems, questioning the benefits of rules as a way to favor firms over markets.

19.1 Introduction

Research into the management of inter firm exchange has, both theoretically and empirically, been dominated by transaction cost theory. The fundamental question addressed by transaction cost theory is why business firms exist (Williamson 1985; Coase 1937). Transaction cost theory proposes that in some cases markets fail to organize transactions efficiently, for instance when contractual hazards associated with transactions are too high. In such cases, instead of relying on the market to exchange

Gerrit Rooks, Chris Snijders, Eindhoven University of Technology
Frits Tazelaar, Utrecht University

https://doi.org/10.1515/9783110647495-019

goods and services, a hierarchy of contractual employee relations is the more effective and efficient option (in-house production). This "make-or-buy" decision is at the heart of the transaction cost theory paradigm, and extensions of this principle, that relate to the extent to which a firm should invest in managing the products or services that they decide to buy, have found their way into economics, sociology, and neighboring disciplines (cf. Batenburg et al. 2003; Raub et al. 2007; Rooks et al. 2000; Rooks et al. 2006; Buskens and Raub 2013; Wynstra et al. 2018). While transaction cost theory's predictions have received some empirical support, many questions remain (cf. Rindfleisch and Heide 1997; David and Han 2004; Geyskens, Steenkamp, and Kumar 2006). One fundamental unexplored area is the structure and content of the hierarchy itself. A hierarchy can be defined as a governance mode that "by means of an authority structure, provides one exchange partner with the ability to develop rules, give instructions, and in effect impose decisions on the others." (Geyskens, Steenkamp, and Kumar 2006: 525).

Under which conditions firms develop rules, and how exchange performance is affected by these rules, are questions that have received comparatively little attention. They are important questions though, since more insight in the costs and benefits of the use of rules will allow researchers to better assess when hierarchies outperform other governance modes such as markets and hybrid governance forms that include relational elements (March et al., 2000). Although formal organizational rules have since long been an important ingredient in theories about bureaucratization of organizations and societies as a whole (Gouldner 1954; Weber 1947), we know surprisingly little about their effects in general (Beck and Kieser, 2003), and their effects on contractual planning in particular. Williamson has stressed on several occasions (Swedberg 1990; Williamson 1996), that the strengths and weaknesses of hierarchy need to be better understood. In his words, economists "are greatly in need of a more adequate theory of bureaucracy", thereby urging sociologists "to call these issues to our attention, lay out their significance, and impress upon us to be responsive" (Swedberg 1990: 124). He also stressed that "in comparison with the study of *market failure*, the study of *hierarchical failure* is seriously underdeveloped. If, however, each generic mode of governance enjoys distinctive strengths and weaknesses, then that disparity should be redressed. Chief among the issues that warrant study in this connection is that of bureaucracy" (Williamson 1996: 17).

The aim of our study is to extend the literature and empirically investigate what the effects of organizational rules are. We explicate and test how (formal) organizational rules – in our case internal organizational rules on procurement – affect the contractual planning of a purchasing transaction. The contractual planning of a transaction is a complex set of activities that comprises negotiating and writing contracts, as well as enforcing contracts. Additionally, we test whether rules affect exchange performance, which we define operationally as the number of problems that occur in the ex post phase of a transaction, and a more subjective measure, namely the satisfaction of the buyer with the product and supplier. We make use of a comprehensive data set containing detailed information about 832 purchasing transactions in Germany.

19.2 Theory and hypotheses

Organizational rules provide procedures for problem solving (Beck and Kieser 2003). We are interested in rules that specify which actions have to be taken by employees or departments given a certain type of transaction. More specifically, we focus on the set of rules developed to deal with issues in the area of procurement and consider how the 'density' of internal organizational procurement rules affects ex ante management of a purchase transaction. Following Schulz (1998: 849) we define rule density as "the number of rules in a given organizational rule population at a given time". Essentially, rule density thus refers to the number of specified rules in a given problem area (as a percentage of the total number of possible rules). In our case the population of rules consists of all potential internal procurement rules. Although it is not obvious that existing rules will necessarily be followed (cf. Borry et al. 2018), we do not consider this issue and assume that, by and large, organizational rules are followed (or at least that the more rules there are, the more rules are followed).

19.2.1 Rules as rational adaptations or irrational coincidences

One of the earliest and most prominent scholars to focus on organizational rules was Max Weber, who explained that the economic advantages of specialization through a division of work can only be realized on the basis of formal rules. Rules and rule-following are an essential element of bureaucracy, which, according to Weber, is a superior form of organization. Formalization increases standardization, precision, and the speed of organizational processes. Formal rules also restrict the organizational members' room for egoistic maneuvers and arbitrary decisions. Moreover, formal rules control and coordinate organizational processes and increase their predictability (Weber 1947). In Weber's view, organizational rules are efficient. One of the reasons for this is that organizational rules store knowledge (Beck and Kieser 2003; Hage 1965; Schulz 1998). Collective experience is transferred to individual and new members through rules 'that reflect but do not reproduce the experiences on which they are based' (March, as cited in Beck and Kieser 2003). The idea that rules are efficient is shared by rational action theories that explain the existence of rules as an efficient way to coordinate action, "a conception of historical efficiency underlies many speculations about rules" (March et al. 2000: 4). Rules are expected to represent what an organization has learned about dealing with matters, both within the organization itself and in relation to other organizations (Williamson 1996).

A more or less opposing view proposes that organizational rules may be inefficient as well. Although it is acknowledged that rules can achieve coordination, in this view rules are often rigid and imprecise means to achieve coordination of activities as well. Blau (1957: 68) expressed this view as follows ". . . effective administration is

contingent on uniform adherence to regulations as well as on adaptability to a variety of specific situations, but bureaucratic pressures compelling strict conformity to rules also give rise to rigidities that interfere with the adaptability needed to handle special cases".

Schultz (1998) distinguished two mechanisms that inhibit or eliminate the effectiveness of rules as organizational learning devices: 'codification traps' and 'sorting'. Codification traps can occur when old rules are applied to new problems: a new situation or problem does not really match a rule, but is used nevertheless. The rule is often 'stretched' or extended to apply better to the problem. Such codification traps are inhibiting organizational learning and lead to inefficiencies because old, inadequate rules might be needlessly reinforced, and the perceived need to create new better rules is decreased. Another inefficiency can be the result of what Schulz calls 'sorting'. Sorting is shorthand for the phenomenon that the most recurring and important problems are encoded into solutions and rules first, while less recurrent and less important problems are encountered and encoded later, or they are not encoded in rules at all. Because of this temporal sequence, rules that are created in response to recurrent problems are likely to be based on relatively old and obsolete knowledge, leading to inefficiencies.

Over time, the discussion on organizational rules has led to literature on closely related concepts such as organizational formalization (the extent to which rules, procedures, instructions, and communications are written), organizational routines (informal rules that came about in an informal way), 'red tape' (rules and routines that do not make sense any more), organizational learning and ecology, and many other, related but subtly different concepts. Whereas theoretical papers about the concepts are plentiful, field studies on organizational rules are relatively rare and often based on archival research (cf. Bozemann and Feeney 2014; Kaufmann and van Witteloostuijn 2018).

19.2.2 Rules and ex ante management

Although there are contradicting ideas about the relationship between organizational rules and contracting, there has been remarkably little large-scale empirical research on effects of rules (Bozemann and Feeney 2014). Above we discussed two arguments that point to the potential inefficiency of organizational rules. In this section we discuss the effects of organizational rules of the buyer on the ex-ante management of a transaction. We focus on the buyer, since this is the empirical focus of the paper. We distinguish two dimension of ex ante management: the content and the effort involved. The content of ex ante management in our case refers to the extensiveness of the contract between buyer and supplier. Second, we discuss effects of organizational rules on the effort involved in the ex-ante management, more specifically, we focus on the cost of contracting, which is measured as the number of full-time equivalent

days spent negotiating and drafting the initial contract. We hypothesize that organizational rules are not necessarily efficient. Organizational rules often lead to more safeguards and higher ex ante transaction costs, than would be necessary on the basis of the relevant characteristics of a transaction by itself.

The first reason is that rules are often founded as a response to crises and serious problems with the aim to prevent that same problem in the future (Zhou 1993), "the stable door is locked after the horse has bolted". Rules are then created to prevent certain problems to reoccur. In situations that resemble the previously problematic situation, preventive measures will now have to be taken because the rules prescribe those preventive actions. Since new situations might well be qualitatively substantially different from the previous problematic situations, the application of the rule is often unnecessary (Schulz 1998). So, rules are applied to prevent problems, and are calibrated (if at all) in such a way that they will be executed also in cases where they would not have been necessary, which obviously leads to higher transaction costs compared to the cases where no or less rules are at play. Rules tend to give descriptions of what to do extra, not what to do less. The latter point also suggests the second argument: as rules typically require additional things to do, there are higher costs involved in terms of time and effort when there are more rules that have to be applied (Ouchi 1979). Furthermore, evidence also suggests that rules are not regularly refreshed or revoked: "[the increase in the number of rules] reflects the creation of many new rules [. . .] without the repeal of rules" (Jakobsen and Mortensen 2016: 302).[1]

Hypothesis 1: The larger the rule-density in a buyer's organization is, the more extensive the contract between the buyer and the supplier will be.

*Hypothesis 2:*The larger the rule-density in a buyer's organization is, the higher the cost of contracting associated with a transaction will be.

19.2.3 Rules and exchange performance: The tradeoff between flexibility and efficiency

In the classical Weberian view, formalization is thought to increase standardization, precision, and the speed of organizational processes. Formal rules also restrict the organizational members' room for egoistic maneuvers and arbitrary decisions, and control and coordinate organizational processes and increase their predictability (Weber 1947). In Weber's view, organizational rules are associated with better

1 A recent paper has also shown that because of the rigor of most rules, employees commonly mention rules as one of the key constraints that they experience in their work (Pindek et al. 2019). Contradicting arguments in the area of organization routines (not rules) can be found in for instance Feldman and Pentland (2003), who argue that routines do not necessarily create inertia and rigor.

performance, and fewer errors. One reason is that organizational rules store knowledge. Formalization codifies best-practice routines to stabilize and diffuse new organizational capabilities (Nelson and Winter 1982). Collective experience is transferred to individual and new members through rules 'that reflect but do not reproduce the experiences on which they are based' (March, as cited in Beck and Kiesler 2003). So, rules are associated with bureaucratic forms of organizations with high levels of standardization, specialization, and ultimately, efficiency.

Formal rules have a downside as well. Too much organizational structure implies that organizations become rigid and inflexible (Sine, Mitsuhashi, and Kirsch 2006). Highly standardized routines make organizations resistant to change (Hannan and Freeman 1984). Rules lock organizations into inflexible patterns of action (Gersick and Hackman 1990). So, according to the literature, structure can have an up and downside. Rules can lead to efficiency and reliability, but also to rigidity and inertia. In line with earlier research, we consider whether there is a tradeoff between efficiency and flexibility (Adler, Goldoftas, and Levine 1999). Too little structure does not guide the required behavior enough, and does not facilitate coordination enough, but too much structure leads to a lack of flexibility and overly cautious interactions. This argument is consistent with the findings in the simulation study by Davis, Eisenhard, and Bingham (2009): organizations with a low or high amount of rules perform worse than those with moderate structure. This leads to the following hypothesis.

Hypothesis 3: Rule density has an inverted U-shaped relationship with exchange performance.

19.2.4 Rule density, environmental dynamism, and exchange performance

There is a long-standing literature that argues that effects of organizational structure on performance depend on the dynamism of the environment of that organization, and the extent to which the rules and routines can keep up with changing circumstances. Contingency theory is the most prominent theory that studies this relation (Galbraith 1973; Schoonhoven 1981). It maintains that there is no one best way of organizing. The effectiveness of a certain organizational form will depend on the degree of task uncertainty. Organizations should adopt a more efficient mechanistic form if the task is simple and stable, while it should adopt a more flexible organic form if the task is complex and changing (Burns and Stalker 1961). A basic assumption that the theory makes is that under conditions of greater uncertainty, more information must be processed to achieve a certain level of performance (Galbraith 1973: 4). As a consequence, organizations are less able to plan ahead or to make decisions under conditions of greater uncertainty, and consequently performance will be worse under such conditions.

In dynamic environments transactions will be more often non-routine tasks. Whereas rules help to efficiently deal with routine purchasing tasks, they are not well

suited for the non-routine purchasing tasks. One way to deal with non-routine tasks is to apply 'old' rules to new circumstances: a new situation or problem does not really match a rule, but is used nevertheless. If rules are applied in completely inappropriate situations, then the contract performance will deteriorate (Schoonhoven 1981). This argument is similar to the "codification traps" idea of Schulz (1998), which also discusses the dangers of applying old rules to new problems. The rule is often stretched or extended to apply better to the problem. Such codification traps lead to inefficiencies, and inhibit organizational learning because old, inadequate rules are reinforced, and the perceived need to create new and better rules is reduced.

A second possible inefficiency because of environmental dynamism can be the result of 'sorting' (Schultz 1998). Sorting refers to the phenomenon that the most recurring problems are encoded into solutions and rules first, while less recurrent problems are encoded later, if at all. Sorting is thus related to the decay of routines and rules as described by Hannan and Freeman (1984). Organizations remember by doing (Nelson and Winter 1982), and hence organizational rules and routines that are not used often, decay. Because of this temporal sequence in sorting, rules that are created in response to recurrent problems are likely to be based on relatively old obsolete knowledge, leading to inefficiencies. Sorting is related to 'routine rigidity', which is a failure to change organizational processes. Routine rigidity may result from the fact that some routines are very tightly aligned with one certain environment, since they are so adapted to this one environment, they are difficult to change because effective and self-reinforcing, and not built to adapt to discontinuities (Gilbert 2005).

Although the argument that environmental dynamism influences the relation between organizational rules and performance is plausible and supported by some empirical results, it has been criticized as being imprecise, since environmental dynamism is a multi-dimensional construct (Davis et al. 2009; Dess and Beard 1984). In their simulation study into the relation between organizational rules and performance Davis et al. (2009) distinguish four dimensions: unpredictability, ambiguity, complexity, and velocity. Velocity and unpredictability refer to global characteristics of opportunities and threats in the environment that are less relevant at the transactional level. Since we focus on the environment of the exchange, and not the organization as a whole, we focus here on performance ambiguity and task complexity. Performance ambiguity refers to a lack of clarity in the exchange because the buyer lacks experience or knowledge to evaluate the partner's performance; task complexity refers to the number of sub-task and the interrelatedness between those sub-tasks.

Hypothesis 3a: The relationship between rule-density and exchange performance is moderated by performance ambiguity. Rules work better when there is less ambiguity.

Hypothesis 3b: The relationship between rule-density and exchange performance is moderated by task complexity. Rules work better when there is less task complexity.

19.2.5 Control Variables: Organizational Capabilities and Transaction Characteristics

More recently, scholars have also recognized that organizational capabilities influence governance decisions (Argyris 1996; Leiblein and Miller 2003; Mayer and Salomon 2006; Mayer 2006; Nickerson and Silverman 2003). Designing a contract has been claimed to be a firm capability that influences contractual planning and, exchange performance in interorganizational relationships. Contract design capabilities reside in the rule within a firm, but also in managers, engineers, and lawyers (Argyris and Mayer 2007). In this paper we control for possible confounding effects of organizational capabilities by including in our models the existence of separate departments with specialized tasks (a legal department, a purchase department, an IT-department, and/or a finance department). These specialized departments form an indication for both governance capabilities and technical capabilities of the firm. Given these capabilities, it seems likely that the ex-ante governance of transactions can take place with higher expertise and lower transaction costs. Although these arguments make it sound as if we assume that the existence of a specialized department is necessarily efficient, many of the considerations about inefficiencies mentioned above with respect to the use of organizational rules may hold for organizational capabilities as well. For example, suppose that the rule is that for large or frequent transactions the legal department should (always) be involved, even if the other contractual hazards (e.g. asset specificity, uncertainty, complexity and/or a lack of competition in the market) are quite low. In such a case, the use of the internal legal specialists is likely to be overkill. However, as with rules, having specialized departments is likely to increase contract density, even without assuming optimal efficient use of the department capabilities.

Other control variables are based on transaction cost theory. Transaction cost theory had its starting point with the seminal article of Coase (1937) 'The Nature of the Firm'. Later, Williamson operationalized the theory by defining particular transaction characteristics that determine governance choices. The most prominent set of transaction characteristics are *specific investments* (e.g., Williamson 1985). If transactions require investments that cannot be redeployed to alternative uses without loss of productive value, then contractual hazards due to dependency arise. Transaction partners can profit opportunistically from this dependency, given that circumstances permit this.

The second transaction characteristic that plays a prominent role in the theory is *uncertainty*, which relates to the inability to predict changes in the environment and partners' behavior. Uncertainty is assumed to vary, meanwhile being always present to at least some extent. Due to uncertainty, combined with bounded rationality, contracting is always costly, and complete contracting not feasible.

A third transaction characteristic that is identified in the theory, although it has received less attention in empirical TCE research (David and Han 2004) is the *frequency* of transactions. Frequency is relevant in two respects: reputation effects and

setup costs (Williamson, 2008: 8). According to Williamson (Williamson 1985: 60) higher levels of transaction frequency provide an incentive for a firm to employ hierarchical governance because "the cost of specialized governance structures will be easier to recover for large transactions of a recurring kind".

Transaction cost theory maintains that there are rational economic reasons for choosing the means of governing transactions. If market transactions become too costly, transactions will be vertically integrated: the firm will choose to arrange matters in-house. In the transaction cost view, an organization exists because it can manage transactions with contractual hazards at lower costs than the market can. Empirically, transaction costs theory has received a fair amount of support, though not always that consistently (Rindfleisch and Heide 1997; David and Han 2004; Geyskens, Steenkamp, and Kumar 2006).

Following Milgrom and Roberts (1992) two other characteristics of the transaction are taken into account as well, namely *task complexity* and (lack of) *competition*. Task complexity precipitates transaction costs by introducing ambiguity about the cause of transaction failure, which in turn makes it difficult to apportion blame between the transacting partners. Moreover, complexity creates a need for coordination among transaction partners. A lack of competition in the supplier's product market may increase the buyer's dependency, and therefore the intensity of competition can be seen as a force that reduces transaction hazards.

19.3 Methods

To test the hypotheses, we make use of a database resulting from a comprehensive survey of IT-purchases by German small and medium sized business (Berger, Kropp, and Voss 2000). This multi-purpose survey was part of a wider Dutch-German cooperative project; the survey was based on an earlier pilot study (Tazelaar, Vaessen, Blumberg, and Raub 1995), as well as on earlier large-scale surveys that were conducted in a related research project in the Netherlands (Batenburg 1997).

The questionnaire contained six major parts, concerning (a) the product and/or service, (b) product and supplier selection, (c) the relationship between the buyer and the supplier, (d) the agreement and contracting, (e) the performance of the supplier, problems, and problem management, and (f) the buyer and his or her relation to the supplier. In total, per transaction more than 300 items were scored. In the survey buyers of information technology in two regions, Halle/Leipzig and Munich, in (then East and West) Germany were sampled. At the time of the survey both regions (Halle/Leipzig and Munich) were economically prosperous regions.

In order for a transaction from a given firm to be included in the survey, the firm had to meet the following requirements: 1) it should be a small or medium sized firm with 4 to 500 employees; 2) it had managed the purchase itself, the decisions about

the product and supplier had been taken by the firm instead of a mother company; 2) an employee of the firm was willing and able to give detailed information about the transaction; 3) the transaction was completed not too long ago – if possible not longer than three years; 4) the transaction should involve only one supplier. To compile the sampling frame the yellow pages were used ('Gelben Seiten für Deutschland. Frühjahr 1999').

The data collection was conducted in two stages. First, a member of the research team contacted the firm in the sample by phone to determine whether the firm met the requirements to be part of the survey, and if this was the case, whether the firm was willing to cooperate in the survey. If a firm agreed to cooperate with the survey, and met the requirements, then a knowledgeable contact person who had been responsible for purchases of information technology was selected and an appointment was made for a face-to-face interview (if firms refused a face-to-face interview a questionnaire was mailed to them).

The telephone interviews started in March 1999 and were concluded in August 1999 (Berger, Kropp and Voss 2002). All firms that agreed to cooperate were sent a letter of confirmation immediately after the telephonic interview. In this letter the selected transaction and the date and time of the appointment were named again. Shortly before the agreed face-to-face interview a member of the research team phoned the firm once more to confirm the appointment. If possible, the respondent was asked after the interview whether he or she was willing to fill out a second (written) questionnaire.

As a result of the high care intensity and the personal assistance in filling out the questionnaires, the response rates to the face-to-face interview were high. The survey team realized 84.2% of all the promised interviews. Additionally, 24.5% of the respondents filled out a second questionnaire. The response to the mail questionnaire was substantially lower (36.4%). The overall response rate was 49%, which is high compared to the standard response rates in organization research (see Kalleberg et al. 1996). The response rate in the region of Halle/Leipzig was higher than in the region of Munich: 57% versus 44% (Rooks, Tazelaar, and Snijders 2010).

The data set that was collected contains detailed information about 1,019 IT-transactions from 832 buying firms. From these 832 buying firms, 645 (= 78%) provided data on a single IT-transaction and 187 (= 22%) provided data on a second IT-transaction as well.

19.4 Measurements

As mentioned before in the data section, the survey was based on a number of earlier large-scale surveys that were conducted in a related research projects in the Netherlands. Many of the measurement items used in this article were tested and calibrated in earlier studies using other data-sets (Rooks and Snijders 2001; Batenburg

et al. 2003; Rooks et al. 2006). Besides our main variables of interest: rule density, firm specialization, contract extensiveness, and ex post performance, we control for several TCE-characteristics and other contextual characteristics.

19.4.1 Organizational characteristics

19.4.1.1 Rule density

To measure rule density, we constructed an instrument that consisted of six items that covered procurement rules. More specifically, the six items were about the management of six essential purchasing tasks, such as the search and selection of suppliers. The items were tested in two small scale survey studies. Every item involved a certain broader problem area, such as which departments or which functions are to be involved in the search, screening, and selection of suppliers. Respondents could indicate whether there were written rules (documents) that covered the problem area. The items and descriptive statistics are shown in Table 19.1.

Table 19.1: Rules and regulations, ordered from most used to least used.

		Nr of obervations	Proportion that has rule	Mokken's H item score
1	Type of agreement / contract that has to be used (an order / standard contract / tailor made contract)	845	0.24	0.69
2	Tendering / collecting and judging tenders or offers	846	0.22	0.76
3	Departments and/or capabilities that are to be involved in search, screening and selection of buyers	848	0.18	0.69
4	Departments and/or capabilities that are to be involved in negotiating, designing and concluding contacts	841	0.16	0.67
5	Departments and / or capabilities that are to be involved in conflict resolution, arbitration, litigation etc.	841	0.14	0.72

Table 19.1 (continued)

	Nr of obervations	Proportion that has rule	Mokken's H item score
6 Evaluating suppliers performance	841	0.13	0.70
7 Benchmarking / periodical supplier evaluations	841	0.13	0.70
8 Supplier audits and product presentations by suppliers	844	0.10	0.59

Rules are often codified in a temporal order. First, the most severe and recurrent problems are codified, whereas less recurrent problems will be codified later (Schulz 1998). To investigate patterns in organizational rules and the scalability of our items we performed a nonparametric item response analysis, the so-called Mokken model (Mokken 1971). This model can be viewed as a probabilistic version of Guttman scale analysis for dichotomous items. An advantage of this nonparametric model compared to parametric versions of item response models such as the well-known Rasch model is that it is more flexible, and sometimes avoids misleading results obtained by parametric models (Junker and Sijtsma 2001; Meijer and Baneke 2004). We used the MSP5 program (Molenaar and Sijtsma 2000) to estimate the model. The results are shown in Table 19.1. According to (Mokken 1971: 185) for practical test construction purposes H values lower than .3 are not scalable, $.3 \leq H \leq .4$ denote a weak scale, $.4 \leq H \leq .5$ denote a medium scale, while H-values above .5 denote a strong scale.

According to the criteria, the items form a very strong scale (Mokken's H-coefficient = 0.68). The reliability of the scale is comparable to Cronbach's alpha and in our case is adequate (Rho = .82). The sum of all eight the items form the variable *rule density*. The finding that the items constitute a Mokken scale is interesting. Apparently, there is an order in the kinds of rules that a firm uses. There obviously exist more frequently used rules, such as rules with respect to the type of agreement that is supposed to be used, and less frequent ones, such as rules about supplier audits. The key finding is that there is a hierarchy in this set of rules, where firms typically tend to implement rules in a bottom-to-top step-by-step way: firms differ in the extent to which they are formalized not only by the number of rules they have, but they differ in how far up the rules-pyramid they are.

19.4.1.2 Organizational capabilities

We measure organizational capabilities using information that the survey respondents provided about the departments in the organization. *Purchasing* is a dummy variable indicating whether or not there was a purchase department present in the firm (18% of

the SMEs had a purchasing department). *Automation* is a dummy variable indicating whether or not there was a automation department (17% of the SMEs had an automation department). *Legal* is a dummy variable indicating whether or not there was a legal department (only 4% of the SMEs had a legal department). *Finance* is a dummy variable indicating whether or not a financial department was present in the firm (21% of the SMEs had a financial department).

19.4.1.3 Firm size

According to Nooteboom, Zwart, and Bijmolt (1992) smaller firms are (more) bounded in their rationality. It is more difficult for smaller firms compared to larger firms to monitor suppliers' performance for instance. In other words, Nooteboom et al. (1992: 144) argue that transaction costs are systematically higher for smaller firms. "Costs of governance schemes to reduce transaction costs are often relatively higher for small firms: transactions may be too small to be worth the bother of such a scheme.". The arguments about firm size related to well-known argument about differentiation and firm size (Blau 1970). We use the number of employees of the buyer at the time of the transaction as an indicator for firm size. The variable *firm size* is the logarithm of the number of employees, with 30 cases (2.94%) imputed using information about sales volume. Likewise, the firm size of the supplier, *size partner*, is the logarithm of the number of employees of the supplier at the time of the transaction; 17 cases (1.67%) were imputed using information about the type of supplier.

19.4.2 Transaction characteristics

19.4.2.1 Asset specificity

Asset specificity reflects the possible exposure to ex post opportunistic holdup caused by specific investments in physical and/or human assets that have little or no value outside of the transaction. We measure asset specificity using four survey questions about switching costs. Switching costs are the expected costs should the supplier be replaced with another supplier. The first question relates to the tendering costs, the costs of searching, screening and selecting a new supplier for a new product. The second question involves the loss in terms of time and money associated (re)training and instruction of personnel. The third question relates to the costs of (renewed) data entry. The fourth question is about the loss in terms of time and money associated with idle production. Together, these four items form a reliable scale (Cronbach's alpha = 0.82).

19.4.2.2 Uncertainty

We measure uncertainty using two survey questions as indicators. The first question relates to the difficulty of assessing a supplier's quality at the time of delivery. The second question is in an historical reference frame; respondents were asked to recall the period of time when the contract was initially signed. The question relates to difficulties in comparing offers between suppliers. The reliability coefficient of this 2-item scale (Cronbach's alpha = 0.73).

19.4.2.3 Transaction size

Transaction frequency is a transaction characteristic that has received scant attention in transaction cost theory, this in sharp contrast to asset specificity and uncertainty. According to Williamson (1985: 60) higher levels of transaction frequency provide an incentive for firm to employ hierarchical governance because "the cost of specialized governance structures will be easier to recover for large transactions of a recurring kind.". The nature of transactions that we study is such that transactions are unlikely to be repeated in substance for any pair of transaction partners. Nonetheless, large projects are more significant determinants of current and future profits of both partners, and hence governance will be more beneficial. We use as a single indicator for transaction size the approximate price (in German Marks). We imputed 46 cases (4.51%) using information about the perceived importance and durability of the product or service. Since the distribution of this variable is highly skewed with many outliers a logarithmic transformation is applied.

19.4.2.4 Task complexity

Task complexity creates a need for coordination between transacting firms. Transactions that involve many different parts, software and hardware that interact in unpredictable way to produce services or products are more complex. We use a measure that was based on two indicators. The first indicator measures the scope of the products and services covered by the transaction. It is a count of the number of products and services covered by the transaction (the questionnaire contained a checklist of 18 components). A second indicator is a categorization of technological complexity developed by Anderson and Dekker (2005) that represents increasing demands for communication and coordination between the buyer and supplier. The correlation between the two indicators is high (r = 0.61). The two indicators form a reliable scale (Cronbach's alpha = 0.69).

19.4.2.5 Competition

Market forces are thought to reduce transaction hazards. Two questions are used to measure competition. The first question asks about the size of the pool of potential suppliers that were identified when the firm searched for suitable transaction partners. The second question asks for the number of different products, i.e. alternatives that could have met the buyer's needs at the time of purchase. The two question correlate highly (r = 0.77) and constitute a reliable scale (Cronbach's alpha = 0.87).

19.4.3 Dependent variables

19.4.3.1 Contract extensiveness

The questionnaire contained a list of 22 financial, legal, and operational issues that can be included in a written (ICT) contract. The respondents were asked to indicate which of the items were included in the written contract. Following Anderson and Dekker (2005), we measure contract extensiveness as the number of items included in the written contract.

19.4.3.2 Cost of contracting

We measure the cost of contracting as the number of full-time equivalent days spent negotiating and drafting the initial contract. A log-transformation is applied to the skewed raw variable (ranges from less than one day to 60). Although the measure has some limitations, for instance it only takes into account costs of contracting that are directly related to labor, it is an improvement over earlier measures (Anderson and Dekker 2005).

19.4.3.3 Ex post transaction problems

A separate section of the questionnaire contained questions about the ex post phase of the transaction. Questions were asked about 11 typical problematic issues that are often associated with IT-transactions (Riesewijk and Warmerdam 1988). Respondents could indicate for each issue to what extent it had occurred and how serious the problem had been. The eleven issues are measured on a five-point scale (ranging from 'not whatsoever' to 'very severe'). The variable *ex post problems* is derived as a scale score on these eleven issues (Cronbach's $\alpha = 0.92$; higher represent more and/or more serious problems). The average score was 14.7. For 282 transactions of the 878 in the sample (32%) a modest or severe problem (3–4) occurred. Most correlations

between the different types of problems are rather high (lowest $r = 0.35$, highest $r = 0.77$), hence problems often appear simultaneously.

19.4.3.4 Exchange performance: Satisfaction buyer

We asked the respondents to give two marks to indicate their satisfaction with the supplier and the product. The respondent could indicate on a scale from 1 ("very good") to 6 ("unsatisfactory"). A third indicator was a question whether the respondent would recommend the supplier "Would you and your employees recommend the supplier to other firms as a result of the delivery of this product, or would you not recommend supplier?" The answer scale ranged from 1 "definitely not recommend" to 5 "definitely recommend". Based on these three questions we constructed a variable *Performance* (Cronbach's $\alpha = 0.83$; higher score represents better performance).

19.5 Results

We estimate our models using ordinary multiple regression analysis, using robust standard errors to account for potential heteroscedasticity, and the partially nested structure of the data (182 firms provided information about two transactions).

Looking at Table 19.2 from a distance, we note several interesting observations. First, transaction characteristics have a strong influence on both contracting characteristics and ex post problems. Higher asset specificity, higher uncertainty, higher complexity, and higher task complexity go with more extensive and more costly contracts. Moreover, contract extensiveness indeed decreases ex post problems (cutting away about half the effect of transaction size, for instance), but even after controlling for contract extensiveness, more complicated transactions (uncertain, large, and complex) are associated with more ex post problems. Another interesting finding is that the only characteristics that seem to matter for exchange performance are uncertainty and rule density (and their interaction).

Focusing on our hypotheses, we see that contracts get more extensive, the more written procedures there exist in the focal firm. This is in concordance with Hypothesis 1. Rules and procedures indeed tend to be associated with more extensive contracts. The data hence support the idea that rules and procedures are of the kind "make sure to perform A and B", and not of the kind "under such and such conditions, you need not worry about C and D". Holding constant transaction characteristics, the more written rules there exist in the focal firm, the higher the cost of contracting. This is in concordance with Hypothesis 2. That is, transactions with a higher 'problem potential' receive more transaction management.

Table 19.2: Multiple regression analyses with robust standard errors on the costs of contracting, contract extensiveness, ex post transaction problems, and exchange performance.

	Cost of contracting	Contract extensiveness	Ex post transaction Problems	Ex post transaction problems	Exchange performance	Exchange performance
Rules						
Rule density	0.12***	0.10***	0.09**	0.08**	−0.09**	−0.08*
Rule density × Rule density				−0.06		0.02
Rule density × Uncertainty				0.13***		−0.12***
Rule density × Complexity				−0.05		0.06*
Separate departments available and firm size						
Purchasing department	0.00	0.09**	0.04	0.04	0.00	0.00
Automation department	0.05	−0.02	0.05	0.05	−0.01	−0.01
Legal department	0.02	0.04	−0.06*	−0.06	0.03	0.02
Finance department	−0.04	−0.04	0.02	0.02	−0.01	−0.01
Firm size	0.02	0.01	−0.08*	−0.07*	0.10**	0.09**
Transaction characteristics						
Size partner	0.02	0.15***	0.08*	0.09***	−0.05	−0.06*
Asset specificity	0.12***	0.24***	0.02	0.03	−0.04	−0.04
Uncertainty	0.11***	0.11***	0.26***	0.27***	−0.20***	−0.21***
Transaction size	0.25***	0.21***	0.13***	0.12***	−0.05	−0.04
Task complexity	0.13***	−0.03	0.11***	0.11***	0.04	0.03
Market						
Competition	−0.09***	0.09***	−0.12***	−0.11***	0.03	0.02
Transaction management						
Contract extensiveness			−0.11***	−0.10***	0.03	0.03
N	877	933	912	912	933	933
F	20.04***	24.19***	12.22***	10.35***	3.72***	3.60***
R-squared	0.26	0.23	0.20	0.22	0.06	0.06

We do not find that there is an optimal number of rules that balances flexibility and efficiency. The quadratic effect of rules finds no support, and alternative ways to test for the inverse U-shape, such as breaking up the analysis in two separate parts (not reported here), also do not show any evidence to support Hypothesis 3. By and large, rules lead to more transaction management, but not to a decrease in problems or an increase in performance. Hypothesis 3 is therefore not supported. We do see that rules are even less effective when there is more performance ambiguity, which supports Hypothesis 3a, albeit in the sense that rules are *less bad* when there is low performance ambiguity. However, this effect is not found for more complex transactions. Hypothesis 3b is not supported.

19.6 Discussion and conclusion

In this study we tested how formal organizational rules affect the contracting costs, contract extensiveness, and exchange performance, namely the number of problems that occur in the ex post phase of a transaction, and overall satisfaction of the buyer. The picture that emerges from our results is one of hierarchical failure. The more organizational rules, the more extensive contracts and contracting costs. The extra management effort that is taken because of existence of rules is not compensated for by better exchange performance, however. To the best of our knowledge this study is the first large scale study into effects of bureaucratization on transaction costs.

Our results have important theoretical implications. A main implication is that the costs of ex ante management and ex-post exchange performance are affected by the level of bureaucracy, especially when a firm is confronted with uncertainty. Bureaucratic costs have been recognized as a factor Williamson (1991: 279) "One advantage of hierarchy over the hybrid with respect to bilateral adaptation is that internal contracts can be more incomplete. [. . .] The advantages of hierarchy over hybrid in adaptation C respects are not, however, realized without cost. Weaker incentive intensity (greater bureaucratic costs) attend the move from hybrid to hierarchy, ceteris paribus." Our results suggest that those bureaucratic costs should be taken into account when studying comparative economic organization. Including bureaucratic costs in the calculus of economic governance of governance, may explain why hybrid organizations are so common.

A second implication is that our findings may offer at least a partial answer to the question that was raised in the meta-analysis of Geyskens, Steenkamp and Kumar (2006). The effect of relational governance on exchange performance was found to be substantially larger than that of hierarchical governance. As an explanation the researchers point to the strength of relational governance, such as flexibility and superior information sharing, but ignore potential weaknesses of hierarchical governance forms. Our results suggest that when hierarchical governance is too

bureaucratic (consists of mainly if-then rules), hierarchies will fail, and hybrid forms of governance are more likely to outperform bureaucratic governance structures.

It is an open question why and when the existence of rules and regulations lead to additional problems. One reason might be that they tend to add detrimental content to a contract. Or, perhaps the rules are such that they lead to other behavior that does not find its way in the contract, but nevertheless leads to more problems than one would expect without the rules. On the other hand, it seems that similar arguments cannot (or need not) be made about purchasing departments: when they are around the contracts get more extensive and that helps preventing problems. However, one might still wonder whether the benefits outweigh the costs in this case. Is a purchasing department worth the money? Our data cannot provide any definitive test on this issue as it needs (even) more detailed measurement.

A second issue that we left untouched is the fact that it might make quite a difference whether the contract itself was made by the buyer or the supplier. The argument for that dates back to at least Macaulay's "battle of the forms": the party who can write down the rules is usually better off. In fact, our data do allow a more thorough test of the importance of who the designer of the contract is, which we leave to a future paper.

A third issue that we did not consider in this study is the genesis of organizational purchasing rules. Under which conditions do firms develop rules? One possibility is suggested by transaction cost theory itself. Empirically, the transaction characteristics asset specificity and uncertainty have received the bulk of research attention. The transaction cost theoretical framework includes a third characteristic, "transaction frequency". Transaction frequency refers to the extent to which transactions recur (Williamson 1985). Williamson suggests that transaction frequency will affect hierarchical governance: "The cost of specialized [i.e., hierarchical] governance structures will be easier to recover for large transactions of a recurring kind. Hence the frequency of transactions is a relevant dimension. Where frequency is low but the needs for nuanced governance are great, the possibility of aggregating the demands of similar but independent transactions is suggested." (Williamson 1985: 60) Then again, one might wonder to what extent rules indeed are such rational adaptations, given that our results suggest that they might be less optimal than one would expect. Perhaps the explanation lies in humans' systematic irrational responses to rational demands instead.

Bibliography

Adler, Paul S., Barbara Goldoftas, and David I. Levine. "Flexibility versus Efficiency? A Case Study of Model Changeovers in the Toyota Production System."; *Organization Science* 10, no. 1 (1999): 43–68.

Anderson, Shannon W., and Henri C. Dekker. "Management Control for Market Transactions: The Relation between Transaction Characteristics, Incomplete Contract Design, and Subsequent Performance." *Management Science* 51, no. 12 (2005): 1734–1752.

Argyres, Nicholas. "Evidence on the Role of Firm Capabilities in Vertical Integration Decisions." *Strategic Management Journal* 17, no. 2 (1996): 129–150.

Argyres, Nicholas, and Kyle J. Mayer. "Contract Design Capabilities and Contract Performance by High Technology Firms: Implications for the Roles of Lawyers, Managers, and Engineers." In *Proceedings of the 8th Annual ISNIE Conference, Tucson Arizona*. 2004.

Argyres, Nicholas, and Kyle J. Mayer. "Contract design as a firm capability: An integration of learning and transaction cost perspectives." *Academy of Management Review* 32, no.4 (2007): 1060–1077.

Batenburg, Ronald S., and A. Van de Rijt. "The External Management of Automation 1995: Codebook of MAT95." *ISCORE paper* 58 (1997).

Batenburg, Ronald S., Werner Raub, and Chris Snijders. "Contacts and contracts: dyadic embeddedness and the contractual behavior of firms." *Research in the Sociology of Organizations* 20, no. 1 (2003): 135–188.

Beck, Nikolaus, and Alfred Kieser. "The Complexity of Rule Systems, Experience and Organizational Learning." *Organization Studies* 24, no. 5 (2003): 793–814.

Berger, Roger, Per Kropp, and Thomas Voss. "Das Management des EDV-Einkaufs 1999." *Codebook. Leipzig: Arbeitsbericht des Instituts für Soziologie* 14 (2000).

Blau, Peter M. "Formal Organization: Dimensions of Analysis." *American Journal of Sociology* 63, no. 1 (1957): 58–69.

Blau, Peter M. "A Formal Theory of Differentiation in Organizations." *American Sociological Review* 35 (1970): 201–218.

Borry, Erin L., Leisha DeHart Davis, Wesley Kaufmann, Cullen C. Merritt, Zachary Mohr, and Lars Tummers. "Formalization and Consistency Heighten Organizational Rule Following: Experimental and Survey Evidence." *Public Administration* 96, no. 2 (2018): 368–385.

Bozeman, Barry, and Mary K. Feeney. *Rules and Red Tape: A Prism for Public Administration Theory and Research: A Prism for Public Administration Theory and Research*. Armonk, NY: M. E. Sharpe 2014.

Burns, Tom, and George M. Stalker. "The Management of Innovation. London." *Tavistock Publishing. Cited in Hurley, RF and Hult, GTM (1998). Innovation, Market Orientation, and Organisational Learning: An Integration and Empirical Examination. Journal of Marketing* 62 (1961): 42–54.

Buskens, Vincent, and Werner Raub. "Rational Choice Research on Social Dilemmas: Embeddedness Effects on Trust." *Handbook of Rational Choice Social Research* (Stanford University Press, Redwood City, CA), Rafael Wittek, Tom A.B. Snijders, Victor Nee, eds. (2013): 113–150.

Coase, Ronald Harry. "The Nature of the Firm." *Economica* 4, no. 16 (1937): 386–405.

David, Robert J., and Shin-Kap Han. "A Systematic Assessment of the Empirical Support for Transaction Cost Economics." *Strategic Management Journal* 25, no. 1 (2004): 39–58.

Davis, Jason P., Kathleen M. Eisenhardt, and Christopher B. Bingham. "Optimal structure, Market Dynamism, and the Strategy of Simple Rules." *Administrative Science Quarterly* 54, no. 3 (2009): 413–452.

Dess, Gregory G., and Donald W. Beard. "Dimensions of Organizational Task Environments." *Administrative Science Quarterly* 29 (1984): 52–73.

Feldman, Martha S., and Brian T. Pentland. "Reconceptualizing Organizational Routines as a Source of Flexibility and Change." *Administrative Science Quarterly* 48, no. 1 (2003): 94–118.

Galbraith, Jay. *Designing Complex Organizations*. Addison-Wesley Longman Publishing Co., Inc., (1973): 156–177.

Gersick, Connie JG, and J. Richard Hackman. "Habitual Routines in Task-Performing Groups." *Organizational Behavior and Human Decision Processes* 47, no. 1 (1990): 65–97.

Geyskens, Inge, Jan-Benedict EM Steenkamp, and Nirmalya Kumar. "Make, Buy, or Ally: A Transaction Cost Theory Meta-Analysis." *Academy of Management Journal* 49, no. 3 (2006): 519–543.

Gilbert, Clark G. "Unbundling the Structure of Inertia: Resource versus Routine Rigidity." *Academy of Management Journal* 48, no. 5 (2005): 741–763.

Gouldner, Alvin W. *Patterns of Industrial Bureaucracy*. New York: The Free Press (1954).

Hage, Jerald. "An Axiomatic Theory of Organizations." *Administrative Science Quarterly* 10 (1965): 289–320.

Hannan, Michael T., and John Freeman. "Structural Inertia and Organizational Change." *American Sociological Review* 49 (1984): 149–164.

Jakobsen, Mads LF, and Peter B. Mortensen. "Rules and the doctrine of performance Management." *Public Administration Review* 76, no. 2 (2016): 302–312.

Junker, Brian W., and Klaas Sijtsma. "Nonparametric Item Response Theory in Action: An Overview of the Special Issue." *Applied Psychological Measurement* 25, no. 3 (2001): 211–220.

Kale, Prashant, Jeffrey H. Dyer, and Harbir Singh. "Alliance Capability, Stock market Response, and Long-Term Alliance Success: The Role of the Alliance Function." *Strategic Management Journal* 23, no. 8 (2002): 747–767.

Kalleberg, Arne L., et al., eds. *Organizations in America: Analysing their structures and human resource practices*. Sage, 1996.

Kaufmann, Wesley, and Arjen van Witteloostuijn. "Do Rules Breed Rules? Vertical Rule-Making Cascades at the Supranational, National, and Organizational Level." *International Public Management Journal* 21, no. 4 (2018): 650–676.

Leiblein, Michael J., and Douglas J. Miller. "An Empirical Examination of Transaction-and Firm-Level Influences on the Vertical Boundaries of the Firm." *Strategic Management Journal* 24, no. 9 (2003): 839–859.

March, James G., Martin Schulz, and Xueguang Zhou. *The Dynamics of Rules: Change in Written Organizational Codes*. Stanford University Press, 2000.

Mayer, Kyle J. "Spillovers and Governance: An Analysis of Knowledge and Reputational Spillovers in Information Technology." *Academy of Management Journal* 49, no. 1 (2006): 69–84.

Mayer, Kyle J., and Robert M. Salomon. "Capabilities, Contractual Hazards, and Governance: Integrating Resource-Based and Transaction Cost Perspectives." *Academy of Management Journal* 49, no. 5 (2006): 942–959.

Meijer, Rob R., and Joost J. Baneke. "Analyzing Psychopathology Items: a Case for Nonparametric Item Response Theory Modeling." *Psychological methods* 9, no. 3 (2004): 354.

Roberts, John, and Paul Milgrom. *Economics, Organization and Management*. Englewood Cliffs, NJ: Prentice-Hall, 1992.

Mokken, Robert Jan "A theory and procedure of scale analysis. The Hague, The Netherlands: Mouton 1971.

Mokken, Robert Jan. *A Theory and Procedure of Scale Analysis: With Applications in Political Research*. Vol. 1. Berlin, Germany: Walter de Gruyter, 2011.

Molenaar, Ivo W., and Klaas Sijtsma. "User's Manual MSP5 for Windows." *Groningen: iecProGAMMA* (2000).

Nelson, Richard R. *An Evolutionary Theory of Economic Change*. Harvard University Press, 2009.

Nelson, Richard R., and Sidney G. Winter. "An evolutionary theory of economic change." Cambridge, Mass. and London, Belknap Harvard, 1982.

Nickerson, Jack A., and Brian S. Silverman. "Why Firms Want to Organize Efficiently and What Keeps Them From Doing So: Inappropriate Governance, Performance, and Adaptation in a Deregulated Industry." *Administrative Science Quarterly* 48, no. 3 (2003): 433–465.

Nooteboom, Bart, Peter Zwart, and Tammo Bijmolt. "Transaction Costs and Standardisation in Professional Services to Small Business." *Small Business Economics* 4, no. 2 (1992): 141–151.
Ouchi, William G. "A Conceptual Framework for the Design of Organizational Control Mechanisms." *Management Science* 25, no. 9 (1979): 833–848.
Pindek, Shani, David J. Howard, Alexandra Krajcevska, and Paul E. Spector. "Organizational Constraints and Performance: an Indirect Effects Model." *Journal of Managerial Psychology* 34, no. 2 (2019): 79–95.
Raub, Werner, Gerrit Rooks, and Frits Tazelaar. "Erträge des Sozialkapitals in Zwischenbetrieblichen Beziehungen: Eine Empirisch-theoretische Studie." *Sozialkapital: Grundlagen und Anwendungen. Sonderheft der KZfSS* 47 (2007): 241–271.
Riesewijk, Bernard., and J. Warmerdam. Het Slagen en Falen van Automatiseringsprojecten [Translation: "Success and failure of IT-projects"]. (1988).
Rindfleisch, Aric, and Jan B. Heide. "Transaction Cost Analysis: Past, Ppresent, and Future Applications." *Journal of Marketing* 61, no. 4 (1997): 30–54.
Rooks, Gerrit, Werner Raub, Robert Selten, and Frits Tazelaar. "How Inter-Firm Co-operation Depends on Social Embeddedness: A Vignette Study." *Acta Sociologica* 43, no. 2 (2000): 123–137.
Rooks, Gerrit, Werner Raub, and Frits Tazelaar. "Ex Post Problems in Buyer–Supplier Transactions: Effects of Transaction Characteristics, Social Embeddedness, and Contractual Governance." *Journal of Management & Governance* 10, no. 3 (2006): 239–276.
Rooks, Gerrit, and Chris Snijders. "The Purchase of Information Technology Products by Dutch SM Es: Problem Resolution." *Journal of Supply Chain Management* 37, no. 3 (2001): 34–42.
Rooks, Gerrit, Frits Tazelaar, and Chris Snijders. "Gossip and Reputation in Business Networks." *European Sociological Review* 27, no. 1 (2010): 90–106.
Schoonhoven, Claudia Bird. "Problems with Contingency Theory: Testing Assumptions Hidden Within the Language of Contingency Theory"." *Administrative Science Quarterly* 26 (1981): 349–377.
Schulz, Martin. "Limits to Bureaucratic Growth: The Density Dependence of Organizational Rule Births." *Administrative Science Quarterly* 43 (1998): 845–876.
Sine, Wesley D., Hitoshi Mitsuhashi, and David A. Kirsch. "Revisiting Burns and Stalker: Formal structure and new venture performance in emerging economic sectors." *Academy of Management Journal* 49, no.1 (2006): 121–132.
Swedberg, Richard. *Economics and Sociology: Redefining their Boundaries: Conversations with Economists and Sociologists*. Princeton University Press, 1990.
Tazelaar, F., P. M. M. Vaessen, B. F. Blumberg, and W. Raub. "Samenwerking tussen Inkoper en Leverancier: Verslag van een Vooronderzoek naar het Management van Inkooptransacties (ISCORE-paper No. 40)." *Utrecht: Universiteit Utrecht* (1995).
Weber, Max. *The Theory of Social and Economic Organization*. Simon and Schuster, 1947.
Williamson, Oliver E. *The Economic Institutions of Capitalism*. New York: The Free Press (1985).
Williamson, Oliver E. "Comparative economic organization: The analysis of discrete structural alternatives." *Administrative Science Quarterly*, 36 (1991): 269–296.
Williamson, Oliver E. *The Mechanisms of Governance*. Oxford University Press, 1996.
Williamson, Oliver E. "Outsourcing: Transaction Cost Economics and Supply Chain Management." *Journal of Supply Chain Management* 44, no. 2 (2008): 5–16.
Wynstra, Finn, Gerrit Rooks, and Chris Snijders. "How is Service Procurement Different from Goods Procurement? Exploring Ex Ante Costs and Ex Post Problems in IT Procurement." *Journal of Purchasing and Supply Management* 24, no. 2 (2018): 83–94.
Zhou, Xueguang. "The Dynamics of Organizational Rules." *American Journal of Sociology* 98, no. 5 (1993): 1134–1166.

Ferry Koster

20 Organizational Innovativeness Through Inter-Organizational Ties

Abstract: In order to be innovative, organizations can benefit from having inter-organizational relations. Through these external relations, organizations get access to valuable resources and they have the possibility to learn from other organizations. At the same time, these ties need to be managed to overcome cooperation problems. Prior studies revealed that inter-organizational relations can contribute to an organization's innovativeness in terms of developing new products and services. This chapter addresses three questions that received little attention to date, namely (1) Does collaborating with other organizations on human resource management (HRM) issues contribute to organizational innovation?; (2) Which of these external ties in the HRM domain matter most for organizational innovation?; and (3) Does the quality of these ties explain organizational innovation?

This chapter aims to shed light on these three questions by analyzing data gathered among 732 private firms from the Netherlands. The analyses show that inter-organizational collaborations in the human resource domain contribute to the innovativeness of organizations (both in terms of innovation performance and innovative human resource management). Furthermore, not all HR collaborations contribute to organizational innovation; organizations having ties with business partners and universities and knowledge centers report the highest levels of innovativeness. And, finally, organizational innovation is higher among organizations that indicate that their HR collaborations contribute to the goals of the organization.

20.1 Introduction

It is widely acknowledged that organizations depend on their environment to produce goods and services (Scott and Davis 2007). This general notion is central to theories as diverse as contingency theory, transaction cost economics, and network theories of inter-organizational relations. These theories focus on the question how organizational structures, strategies, and outcomes are affected by characteristics of the environment in which these organizations operate. This means that they belong to the branch of theories that regard organizations as open systems, as opposed to closed system approaches which do not take the organizational environment into account (Scott and Davis 2007). These open system theories dominate the field of organization studies. The dominance of this view is also illustrated by Baum and Rowley (2002: 3) when they state that: "Although historically, rational, natural and open systems definitions have been

Ferry Koster, Department of Public Administration and Sociology, Erasmus University Rotterdam

https://doi.org/10.1515/9783110647495-020

associated with distinct research programs, each with its own conceptual frameworks, guiding assumptions, and empirical approaches, contemporary perspectives built on these foundations invariably take an open systems view, and combine it with either a rational or a natural systems orientation". In other words, all modern theories of organizations belong to the open systems perspective (Scott and Davis 2007).

Network theories focus on a specific part of organizational environments, namely the relationships that organizations have with other actors (organizations, government bodies, customers, and so forth). Research in this field generated insights concerning the conditions under which these relationships are established and how they are sustained (Gulati and Gargiulo 1999; Rooks et al. 2000). Furthermore, research lead to in-depth knowledge of configurations of inter-organizational relations (Pittaway et al. 2004), while other research focused on understanding the structure of inter-organizational networks and aimed at investigating how these ties affect organizational outputs such as financial and innovation performance (e.g. Ahuja 2000; Oerlemans, Meeus, and Boekema 1998; Schilling and Phelps 2007).

Nevertheless, several issues received little attention to date, while the literature suggests that they may matter to understand organizational innovativeness. First, while there is plenty of research concerning how network structures and network positions affect organizational innovation and several studies show that access to external resources explains the degree of organizational innovativeness (Faems, Van Looy, and Debackere 2005), far less is known about *which* actors matter most for the innovation performance of organizations. This calls for research focused on the type of actors with which organizations interact and whether this relates to their innovativeness. In addition to that, research focuses on collaborations between organizations on issues such as product development and production of good and services, but not on collaborations in the domain of human resource management. This latter type of collaboration gained attention with the growing interest in organizational eco-systems (Von Krogh and Geilinger 2014), but how this relates to organizational innovation is unknown. Hence, more should be known about whether *human resource collaboration* matters for organizational innovation. Thirdly, most innovation research focuses on a specific kind of innovation, namely improvements regarding the production of new goods and services (Pouwels and Koster 2017). Some studies also investigate changes in organizational structures and processes, which are also part of the innovation performance of organizations (Maine, Lubik and Garnsey 2012). But, the impact of inter-organizational relations on the *innovativeness of the human resource management* of organizations has not been investigated to date. Instead, explanations of innovative human resource management focused mainly on intra-organizational characteristics and overlooked inter-organizational relations. At the same time, a large body of the human resource management literature argues that external fit – the alignment of human resource practices to the organizational environment – is an essential part of the effective management of people (e.g. Ulrich and Dulebohn 2015). What is more, most of these studies focus on best practices that

are supposed to contribute to a higher performance of organizations (Huselid 1995), whereas innovation of human resource practices is a matter of adoption. Hence, the theoretical notion of fit has a strong foothold in this literature, but is not often empirically investigated. This calls for research connecting the innovativeness of human resource practices of organizations to their external ties.

Based on these observations concerning the current state of research, this study has the following aims, namely (1) to assess whether human resource collaboration matters for organizational innovativeness in general; and (2) to investigate which of these inter-organizational ties matter most for organizational innovativeness. In this study, the focus is on organizational innovativeness in the broad sense, meaning that both the innovation performance as well as innovations in the human resource practices that organizations apply are investigated. Data from a recently conducted survey among 732 organizations in the Netherlands are analyzed to generate insights about the role of resources for innovation performance and innovative human resource management.

20.2 Two types of innovation

Innovation refers both to "creating new things" and "doing things differently" (Maine, Lubik, and Garnsey 2012). While the first conception of innovation received much attention in the literature, the second approach to innovation is far less investigated. As a result, much is known about the creation of novel outcomes by organizations. However, there are good reasons to assume that innovation reflects a broader strategy of organizations that also includes exploring new markets, renewing organizational processes, and so forth (Crossan and Apaydin 2010; Pouwels and Koster 2017). While innovation research emphasizes that organizations can improve organizational processes in different domains, this literature remains largely separated from the literature on innovativeness with regard to functional fields, in particular with regard to the introduction of new ways of managing employees. In other words, innovation studies and human resource studies have not informed each other. Reviewing the literature on innovative human resource management (HRM), Koster (2019) shows that there are least three different approaches to innovative HRM, namely studies examining the innovativeness of human resource practices and policies, HRM innovativeness in response to external developments, and studies linking HRM policies and practices to the innovation performance of organizations. Following this threefold distinction, the present study fits the second strand of the literature in which HRM innovativeness is linked to the external environment.

While several authors state that having ties with other organizations, granting access to their resources and knowledge, is a key ingredient for organizational innovation, research mostly focuses on intra-organizational explanations of innovativeness. In an

extensive overview of the literature, Crossan and Apaydin (2010) show that most research investigates the role of micro level factors such as individual creativity and team structures and organizational factors such as organizational structure, complexity, and slack. At the macro level, the focus is on industry structures and innovation systems (Crossan and Apaydin 2010). Hence, compared to other explanations of innovation, the number of studies investigating the impact of resources accessed through external networks – the meso level – remains a field to be developed. Based on the premise that organizational learning and information sharing are vital to knowledge economies (Adler 2001), such external ties can contribute to organizational innovation.

20.3 Sources contributing to innovation

There are contrasting predictions about the relationship between inter-organizational ties and organizational innovation. One the one hand, there are theories emphasizing the risks of cooperation. These risks result from a loss of control (Gnyawali and Park 2009), the occurrence of opportunistic behavior (Van Haverbeeke, Duysters, and Hagedoorn 2002), and difficulties relating to the transfer of knowledge between organizations (Lam 1997). Based on these risks of cooperation, it is argued that inter-organizational collaboration hinders organizational innovation. However, in a recent study, Pouwels and Koster (2017) show that these risks do not dominate inter-organizational collaborations aimed at creating innovations, as they find that the two are positively linked across a sample of European companies. Hence, their results empirically support theories emphasizing the benefits of cooperation. These theories emphasize the importance of having access to external sources contributing to organizational innovation (Nooteboom 1994), risk reducing strategies to enhance cooperation (Hagendoorn 2002), and the transfer of knowledge between organizations (Ahuja 2000). The kinds of collaboration investigated in the present study also reflect such contributions to organizational innovation.

One of the reasons for this finding lies in the management of these external collaborations. First, organizations that manage to create cooperative relations with others, for example because they interact repeatedly and have a sufficiently long shadow of the future to solve trust problems (Buskens and Raub 2002), may thus benefit terms of organizational innovativeness. And, secondly, organizations may be less likely to stay in unproductive or risky collaborations. This is of course not to say that the collaborations are completely free from the risk of cooperation, but at least there seems to be some logic in the argument that they be disbanded as soon as these risks dominate the relation, unless they are forced to collaborate or if organizations their choice in collaboration partners is extremely limited, for example because there are no alternative partners. But again, especially in the latter case, organizations are likely to withdraw from these collaborations. If there is a lack of alternatives, this implies that they will not

collaborate with other organizations. This latter strategy – limiting the dependence on other organizations – can be means of dealing with this problem.

Organizations that manage to create collaborative ties with other organizations can benefit from the advantages that collaborating with others may have in terms of learning and resources being shared among these organizations. Having access to these resources and being able to learn from the experiences of other organizations in turn are condition for organizational innovation (Crossan and Apaydin 2010). Hence, it is expected that in general, collaborating with other organizations contributes to the innovativeness of organizations, as was found in earlier research for organizational collaboration in domains such as product development and marketing (Pouwels and Koster 2017). Here it is argued that the impact of external collaboration can be extended to collaboration on issues related to the human resource management of organization as it offers a means of solving challenges collectively. For example, an organization that needs to train workers may not have the means to do this individually, while it is possible to develop training programs in collaboration with others. Based on these considerations, the first hypothesis is formulated.

Hypothesis 1: There is a positive relationship between organizational innovativeness and HR cooperation.

This first hypothesis states that having ties with other organizations contributes to the innovativeness of organizations. It may, however, be the case that the contributions depend on the type of collaboration partner. Several mechanisms may be at work, depending on the kind of collaboration partner. Collaborating with peers and similar organizations may add less to the organization than ties with dissimilar organizations, because these actors possess little extra knowledge and information from which the organizations can benefit. However, collaborating with dissimilar organizations may be far riskier than having ties with similar organizations, for example because it is more difficult to estimate whether the other organization actually puts in the effort and resources as promised. Next to the argument that novelty of information and access to unique resources can contribute to organizational innovativeness, it can be argued that similar organizations are actually interesting cooperation partners. The argument is that even though both partners may have access to similar information, they can learn from each other's experiences much more easily than dissimilar organizations. It can be assumed that the issues that one organization within a sector or that produces particular goods or services will also be encountered by other organizations in that sector or that produces similar goods and services. As a result, there is added value in having ties with similar organizations. This means that ties with similar organizations are more easily managed, but add less new information and that organizations face more costs to manage ties with dissimilar organizations, while they may also lead to higher returns (Hoffmann and Schlosser 2001; Tansky and Heneman 2003; Van Gils and Zwart 2004). Since similarity is a matter of degree and the relative importance of novelty of the innovation and the costs of managing the external tie are difficult to estimate beforehand, it

is not evident which partner adds the most to the innovativeness of organizations. Hence, the following hypothesis is formulated.

Hypothesis 2: The impact of HR cooperation on organizational innovativeness differs across HR cooperation partners.

The previous hypotheses are based on the existence of a tie with the HR cooperation partners. This part of the analysis does not inform us about the quality of these relationships. While the second hypothesis is based on the assumption that the costs and benefits of ties with other organizations may vary, it does not yet test whether this is the case. More specifically, it is expected that both the costs and benefits of these ties increase if the partner organization is more dissimilar. The costs and benefits translate into the extent to which having a tie with other organizations contribute to the goals of the organizations, the extent to which these ties add value to the organization, the level of complexity associated with having these ties and the uncertainty involved (Bachmann 2003; Beugré and Acar 2008).

Hypothesis 3: Organizational innovativeness is higher if the HR collaborations is viewed as more beneficial and lower if these collaborations are viewed as costly.

20.4 Data and method

20.4.1 Data

Data from the Innovative HRM Survey (Koster et al. 2017) are analyzed to investigate the relationship between resources and organizational innovation. These data were collected among a random sample of Dutch firms using an online questionnaire. The survey includes several characteristics of organizations – such as the composition of the workforce and inter-organizational relations – as well as their level of innovativeness in different domains. The data were collected by Kantar Public using their panel with private organizations (NIPObase Business). This panel consists of 15,000 representatives (owners and human resource managers) from Dutch firms. From this panel, a random selection of 3,000 organizations was drawn. In total 752 firms responded (a response rate of 25 percent). Some variables are not available for all organizations. The final data set consists of 732 organizations. These organizations operate in different economic sectors and differ in size. The dataset includes a large number of small organizations. About 90 percent of the responses are from organizations with 1 to 9 employees and 6 percent of the organizations in the dataset have 10–50 employees. Hence, the dataset takes into account that most organizations in the Netherlands have less than 10 employees (about 96 percent according to Statistics Netherlands) and that about 3 percent of the organizations have 10–50 employees.

20.4.2 Measures

20.4.2.1 Dependent variables: Innovation performance and innovative HRM

Innovation performance

To get a broad measure of innovation performance (one that goes beyond single item measures of product and service innovation) it is asked whether the organization (1) Developed goods or services that are new for his organization (but already available on the market); (2) Introduced goods or services that were not on the market yet; (3) Strongly improved existing goods and services; (4) Introduced new ways of marketing goods and services; and (5) Introduced new organizational processes. Together, these items reflect several aspects of organizational innovation, which were already discussed by Schumpeter in 1934 who argued that organizational innovation involves the introduction of new goods and services, as well as finding new markets and the need to adapt organizational processes. This measure relies on earlier operationalizations, such as the Community Innovation Survey, with the specific aim to get an overall indication of a firm's innovativeness (Armbruster et al. 2008).

Innovative HRM

To assess the extent to which the organizations engage in innovative HRM, a scale was developed based on research on innovative HRM. Innovative HRM is measured with a scale consisting of four questions about whether the organization renewed their human resource function. The exact wording is: "Has your organization renewed . . . ." followed by four statements about the human resource functions, namely "hiring personnel", "outplacement of personnel", "internal mobility of personnel", and "workforce composition". Respondents were asked to indicate how much this applied to their organization on a 5-point scale (running from 1 = does not apply at all to 5 = applies completely). This measure captures the idea of organizational innovation to the domain of human resource management. It closely follows the approach of Agarwala (2003), but with an important difference. While in that study, managers were asked to rate the innovativeness of a list of human resource practices, the present study asks about renewal, which is more in line with studies of innovation performance (Koster and Benda 2020).

A principal component analysis was performed to investigate the structure of the two dependent variables. The results are reported in Table 20.1. We can conclude that the scales measuring innovation performance and innovative HRM indeed differ from each other. The items measuring innovative performance belong to one dimension and the items related to changes in human resource practices of organizations belong to a different dimension. Both scales are internally consistent: the Cronbach's alpha of innovation performance is 0.855 and the Cronbach's alpha of innovative HRM is 0.936.

Table 20.1: Principle component analysis of innovation performance and innovative HRM.

Item	1	2
Innovation performance		
Products and services: new for his organization	**0.780**	0.183
Strongly improved existing products and services	**0.776**	0.126
Products and services: new for the market	**0.775**	0.234
New ways of marketing products and services	**0.774**	0.259
Introducing new organizational processes	**0.734**	0.240
Innovative human resource practices		
Innovations in . . .		
. . . hiring personnel	0.224	**0.896**
. . . outflow of personnel	0.194	**0.885**
. . . workforce composition	0.265	**0.885**
. . . internal mobility of personnel	0.261	**0.881**
Eigen value	3.174	3.375
% explained variance	35.272	37.495
Cronbach's alpha	0.855	0.936

Notes: N = 732 organizations
Varimax rotation. Factor loadings > 0.30 in bold
Source: Innovative HRM Survey

20.4.2.2 Independent variables: HR cooperation and HR cooperation partners

To measure whether organizations collaborate with other organizations, respondents were asked to indicate whether they collaborated with others on four issues related to the management of human resources (such as hiring personnel and the outflow of personnel). It was asked whether they collaborate with the following others: (1) Competitors; (2) Business association partners; (2) Competitors; (3) Suppliers; (4) Buyers; (5) Universities and knowledge centers; and (6) Public organizations. This variable has the value "0" if the answer is "no" and "1" if the organization did collaborate with that partner. As a result, information is available about a specific kind of inter-organizational tie (namely collaboration in the area of human resource management), instead of an overall indication of the ties with other organizations. The downside of having this very specific indicator is that it cannot be ruled out that the organizations collaborate on other issues.

Two independent variables are constructed using these measures. The variable *HR cooperation* is constructed by summing the responses to the six questions. This indicates the number of HR cooperation partners the organization has. The variable *HR cooperation partners* consists of the separate dummy variables.

Tie quality is measured by asking respondents to rate the ties with other organizations regarding the extent to which these ties contribute to the goals of the

organization, adds value, are complex, and how much uncertainty there is surrounding these ties. This resulted in four dummy variables indicating these aspects of the external ties.

20.4.2.3 Control variables

The following control variables are included in the analyses. *Organization size* is measured by asking respondents to indicate the number of employees that the organization has. Prior studies show that the relation between organization size and organizational innovation is curvilinear (an inverted U-shape) (Nitin and Gulati 1996; Heunk 1998; Koster 2018). In line with these prior studies, the *quadratic term of organization size* is also added to the models. *Sector* was measured by asking respondents in what economic sector the organization operated. The variable *permanent employees* was measured by asking respondents to indicate to what extent the organization consists of employees with a permanent contract (measured on a 5-point scale). The variable *highly educated* was measured with a 5-point scale indicating to what extent the organization employs highly educated employees. The variable *firm specific knowledge* was measured with a 5-point scale indicating to what extent firm specific knowledge and skills are important for organizational performance. Table 20.2 provides an overview of the variables included in the analyses.

Table 20.2: Descriptive statistics of the measured included in the analyses.

	Min/Max	Mean	Standard deviation	Percentage
Innovation performance	1/5	2.41	0.95	
Innovative HRM	1/5	1.80	0.91	
Organization size	1/5	1.17	0.58	
Organization size (categories)				
1–9	0/1			89.50
10–49	0/1			6.50
50–99	0/1			1.70
100–249	0/1			0.90
250 or more	0/1			0.90
Sector				
Industry and production	0/1			4.70
Construction	0/1			6.60
Retail – food	0/1			3.10
Retail – nonfood	0/1			13.20
Whole sale	0/1			7.40
Cars and repair	0/1			1.90
Catering	0/1			3.90

Table 20.2 (continued)

	Min/Max	Mean	Standard deviation	Percentage
Transport and communication	0/1			3.20
Business services	0/1			35.20
Other services	0/1			10.20
Information technology	0/1			8.50
Financial institutions	0/1			2.10
Permanent employees	1/5	2.99	1.69	
Higher educated	1/5	2.94	1.58	
Firm specific knowledge	1/5	3.68	1.35	
HR cooperation partners				
Competitors	0/1			11.8
Business association partners	0/1			18.6
Suppliers	0/1			11.7
Buyers	0/1			10.4
Universities and knowledge centers	0/1			10.4
Public organizations	0/1			7.8
Tie quality				
Goals	0/1			28.8
Added value	0/1			17.8
Complexity	0/1			15.7
Uncertainty	0/1			11.1

Notes: N = 732 organizations
Source: Innovative HRM Survey

20.4.3 Method

OLS regression analyses are performed with innovation performance and innovative HRM as the dependent variables. For these two dependent variables, three models are estimated. Each model includes the control variables and the variables measuring whether the organization collaborates with the different partners. In the first model, HR collaboration is added to the models. The second model includes the separate HR collaboration partners. And, in de final model, the quality of the ties with the HR partners is added.

20.5 Results

20.5.1 Descriptive results

Focusing on the variables of interest, Table 20.2 shows the following. First, on average, organizations have a higher score on innovation performance than on innovative

HRM. Secondly, looking at the HR cooperation partners of organizations, it turns out that organizations most often collaborate with organizations belonging to the same business association (18.6 percent of the organizations have such partners) and that public organization are less commonly chosen as HR cooperation partners (7.8 percent of organizations have these partners). This is also shown in Figure 20.1. Furthermore, it should be noted that the majority of organizations (64 percent) does not collaborate with one of these partners. That 36 percent of the organizations collaborates on HR issues can be considered low, compared to other kinds of collaboration. Data from the European Company Survey show that in the Netherlands, 68 percent of the organizations collaborate to design goods and services, 66 percent collaborates to produce goods and services, and 62 percent collaborates with other organizations to sell goods and services (Koster 2015). Most of the organizations that do have HR cooperation partners, only cooperate with one of these partners (17 percent). Of the 732 organizations investigated here, there are 5 that report that they collaborate with all 6 partners on HR issues.

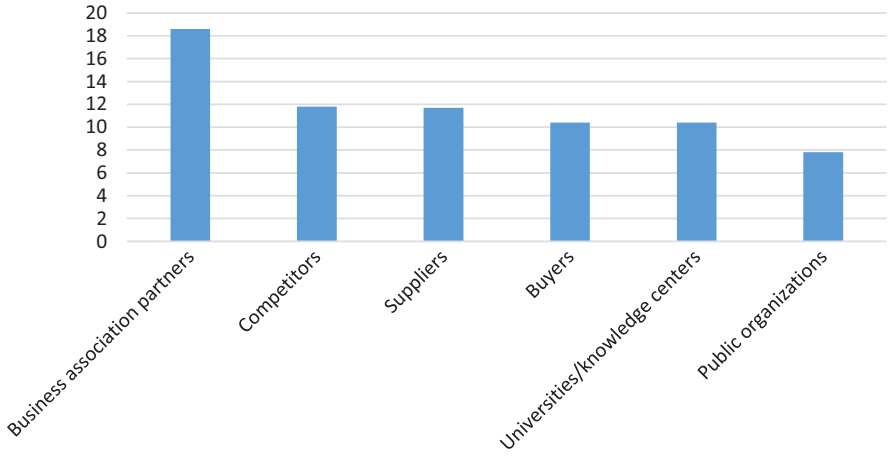


Figure 20.1: HR cooperation partners (% of organizations).
Note: N = 732 organizations
Source: Innovative HRM Survey

20.5.2 Regression analysis

The results of the regression analyses are reported in Table 20.3 (innovation performance) and Table 20.4 (innovative HRM). While there are some notable differences between the models with regard to the control variables (for example, firm-specificity matter for innovation performance but not for innovative HRM), the patterns of the hypotheses are largely the same. Hypothesis 1 is supported: the more

Table 20.3: Regression analysis on innovation performance.

	(1)			(2)			(3)		
Intercept	0.52	*	0.21	0.58	**	0.22	0.51	*	0.22
Number of employees	0.77	**	0.23	0.71	**	0.24	0.81	**	0.24
Number of employees ^2	−0.12	**	0.05	−0.12	*	0.05	−0.12	**	0.05
Sector									
Industry and production	−0.05		0.17	−0.07		0.17	−0.07		0.17
Construction	−0.16		0.11	−0.16		0.11	−0.18		0.11
Retail – food	0.24		0.19	0.27		0.19	0.27		0.19
Retail – nonfood	−0.04		0.11	−0.04		0.11	−0.05		0.11
Wholesale	0.36	**	0.13	0.36	**	0.13	0.33	**	0.13
Cars and repair	−0.18		0.23	−0.21		0.24	−0.19		0.24
Catering	0.06		0.15	0.07		0.15	0.09		0.15
Transport and communication	−0.37	*	0.19	−0.36		0.19	−0.41	*	0.19
Other services	0.09		0.12	0.10		0.12	0.09		0.12
Information technology	0.10		0.13	0.10		0.13	0.08		0.14
Financial institutions	−0.16		0.22	−0.15		0.22	−0.10		0.23
Business services (reference)									
Permanent employees	0.05	*	0.02	0.05	*	0.02	0.04	*	0.02
Higher educated	0.06	**	0.02	0.06	*	0.02	0.07	**	0.02
Firm specific knowledge	0.20	***	0.02	0.20	***	0.02	0.20	***	0.02
HR cooperation	0.18	***	0.03						
HR cooperation partners									
Competitors				0.01		0.11			
Business association partners				0.29	**	0.09			
Suppliers				0.10		0.11			
Buyers				0.27	*	0.12			
Universities and knowledge centers				0.34	**	0.12			
Public organizations				0.06		0.13			
Tie quality									
Goals							0.23	*	0.10
Value							0.14		0.11
Complexity							0.06		0.13
Uncertainty							0.02		0.13
Adjusted R^2	0.27	***		0.27	***		0.25	***	
R^2 change	0.29	***		0.29	***		0.27	***	

Notes: N = 732 organizations
*p < 0.05; **p < 0.01; ***p < 0.001
Source: Innovative HRM Survey

an organization collaborates with other organizations in the HR domain, the more innovative the organization is. This holds for innovation performance and innovative HRM. As expected in hypothesis 2, some of the HR collaborations add more in terms of organizational innovation than others. More specifically, organization collaboration in the HR domain with business association partners, universities and

Table 20.4: Regression analysis on innovative HRM.

	(1)			(2)			(3)		
Intercept	−0.31		0.19	−0.24		0.19	−0.34		0.19
Number of employees	1.52	***	0.20	1.45	***	0.20	1.58	***	0.21
Number of employees ^2	−0.22	***	0.04	−0.21	***	0.04	−0.22	***	0.04
Sector									
Industry and production	0.07		0.15	0.06		0.15	0.07		0.15
Construction	0.11		0.09	0.13		0.09	0.08		0.10
Retail – food	0.38	*	0.17	0.42	*	0.17	0.42	*	0.17
Retail – nonfood	0.07		0.09	0.08		0.09	0.06		0.09
Wholesale	0.29	**	0.11	0.30	**	0.11	0.26	*	0.11
Cars and repair	0.84	***	0.20	0.81	***	0.20	0.81	***	0.20
Catering	0.52	***	0.13	0.54	***	0.13	0.56	***	0.14
Transport and communication	0.18		0.17	0.18		0.17	0.13		0.17
Other services	−0.08		0.11	−0.07		0.11	−0.08		0.11
Information technology	0.06		0.12	0.06		0.12	0.04		0.12
Financial institutions	0.16		0.20	0.16		0.20	0.22		0.20
Business services (reference)									
Permanent employees	0.13	***	0.02	0.12	***	0.02	0.12	***	0.02
Higher educated	0.01		0.02	0.00		0.02	0.01		0.02
Firm specific knowledge	0.01		0.02	0.02		0.02	0.01		0.02
HR cooperation	0.19	***	0.03						
HR cooperation partners									
Competitors				0.10		0.09			
Business association partners				0.32	***	0.08			
Suppliers				0.06		0.10			
Buyers				0.23	*	0.11			
Universities and knowledge centers				0.48	***	0.11			
Public organizations				−0.02		0.12			
Tie quality									
Goals							0.25	**	0.09
Value							0.15		0.10
Complexity							0.17		0.11
Uncertainty							−0.09		0.12
Adjusted R^2	0.41	***		0.41	***		0.39	***	
R^2 change	0.42	***		0.43	***		0.41	***	

Notes: N = 732 organizations
*p < 0.05; **p < 0.01; ***p < 0.001
Source: Innovative HRM Survey

knowledge centers, and to some extent those working with buyers report higher levels of innovation performance and innovative HRM. Having ties with competitors, suppliers, and public organization turn out not to matter for organizational innovation. Finally, with regard to the quality of the ties with these partners, only 1 aspects seems to matter, namely the extent to which these HR collaborations contribute to

organizational goals is positively related to organizational innovation. These models also show that costly collaborations (in the sense that they are viewed as complex or uncertain) do not undermine organizational innovativeness. Hence, hypothesis 3 is only partly supported.

20.6 Conclusions

The analyses presented here show that HR collaborations contribute to the innovativeness of organizations, both in terms of innovation performance and innovative HRM. Furthermore, the results show that the added value of having ties with collaboration partners in terms of organizational innovation differs. In particular, ties with business association partners and universities and knowledge centers contribute to organizational innovativeness. And, finally, the outcomes suggest that HR collaboration partners are relevant for organizational innovativeness if they contribute to the goals of the organization.

Whereas prior studies have focused on the link between inter-organizational relationships and innovation, this study provides several new insights, by extending the scope of the analysis. The insight that organizations collaborating with organizations that belong to the same business association and those collaborating with universities and knowledge centers are more innovative. Whereas previous studies have focused on network structures and diversity in resources, this suggests that it also matters with whom an organization collaborates and what the other side has to offer. It makes sense to make a distinction regarding the basis of the connection between the organizations. Whereas prior studies mainly focused on whether there are ties between organizations or use measures of technical dimensions of collaboration (e.g. whether organizations cooperate on issues such as design and product development), the present study focuses on collaboration on personnel-related issues. By having two indicators of organizational innovation, namely innovation performance and innovative HRM, it is possible to compare the outcomes for these indicators. Overall, the patterns are similar, but since the outcomes are somewhat more pronounced for innovative HRM, seems to suggest that innovations in one domain (in this case renewal of human resource management policies and practices) relate to external ties in the same domain (collaboration on human resource management related issues).

It turns out that ties with public organizations do not matter for organizations, but that collaborating with universities and knowledge centers is related to organizational innovation. As already noted, this may be explained by governance issues that seem to be more complicated if a private organization collaborates with public organizations than with universities and knowledge centers. In times in which there is debate about the added value of institutions such as universities and there is greater

emphasis on generating knowledge applicable knowledge, this outcome shows that these institutions perform this task quite well already.

Finally, this study sheds light on an issue that received little attention in the literature to date, namely HR collaboration. Some of the work that has been conducted in this area remained theoretical in nature and empirical tests are scarce (for example, a theoretical article by Gardner from 2005 on the topic of human resource alliances is still largely untested). With regard to the outcomes of such collaborations, even less is known. The analyses presented here make a case for further investigating the connection between collaborating on HR issues and organizational innovation.

References

Adler, Paul. S. 2001. "Market, Hierarchy, and Trust: The Knowledge Economy and the Future of Capitalism." *Organization Science* 12: 215–234.

Agarwala, Tanuja. 2003. "Innovative Human Resource Practices and Organizational Commitment: An Empirical Investigation." *International Journal of Human Resource Management* 14: 175–197.

Ahuja, Gautam. 2000. "Collaboration Networks, Structural Holes, and Innovation: A Longitudinal Study." *Administrative Science Quarterly* 45: 425–455.

Armbruster, Heidi, Andrea Bikfalvi, Steffen Kinkel, and Gunter Lay. 2008. "Organizational Innovation: The Challenge of Measuring Non-Technical Innovation in Large-Scale Surveys." *Technovation* 28: 644–657.

Bachmann, Reinhard. 2003. "The Coordination of Relations Across Organizational Boundaries." *International Studies of Management & Organization* 33: 7–21.

Baum, Joel A.C. and Tim J. Rowley. 2002. "Companion to Organizations: An Introduction." Pp. 1–34 in *The Blackwell Companion to Organizations*, ed. Joel A.C. Baum. Oxford: Blackwell Publishers.

Beugré, Constant D., and William Acar. 2008. "Offshoring and Cross-Border Interorganizational Relationships: A Justice Model." *Decision Sciences* 39: 445–468.

Buskens, Vincent and Werner Raub. 2002. "Embedded Trust: Control and Learning." Pp. 167–202 in *Advances in Group Processes*, ed. Edward J. Lawler and Shane R. Thye. Amsterdam: JAI/Elsevier.

Crossan, Mary M. and Marina Apaydin. 2010. "A Multi-Dimensional Framework of Organizational Innovation: A Systematic Review of the Literature." *Journal of Management Studies* 47: 1154–1191.

Faems, Dries, Bart Van Looy and Koenraad Debackere. 2005. "Interorganizational Collaboration and Innovation: Toward a Portfolio Approach." *Journal of Product Innovation Management* 22: 238–250.

Gnyawali, Devi and Byung-Jin Park. 2009. "Co-opetition and Technological Innovation in Small and Medium-Sized Enterprises: A Multilevel Conceptual Model." *Journal of Small Business Management* 47: 308–330.

Gulati, Ranjay and Martin Gargiulo. 1999. "Where do Interorganizational Networks Come From?" *American Journal of Sociology* 104: 1439–1493.

Hagedoorn, John. 2002. “Inter-Firm R&D partnerships: An Overview of Major Trends and Patterns Since 1960.” *Research Policy* 31: 477–492.
Heunks, Felix J. 1998. “Innovation, Creativity and Success.” *Small Business Economics* 10: 263–272.
Hoffmann, Werner H. and Roman Schlosser. 2001. “Success Factors of Strategic Alliances in Small and Medium-Sized Enterprises. An Empirical Survey.” *Long Range Planning* 34: 357–381.
Huselid, Mark A. 1995. “The Impact of Human Resource Management Practices on Turnover, Productivity, and Corporate Financial Performance.” *Academy of Management Journal* 38: 635–672.
Koster, Ferry. 2015. *Collaboration and Innovation in SMEs in Europe and the Netherlands. An Exploration Using the European Company Survey*. ICOON Paper 1. Tilburg: ICOON.
Koster, Ferry. 2018. “Personeelsbeleid in de Platformeconomie.” *Mens en Maatschappij* 93: 283–305.
Koster, Ferry. 2019. “Innovative HRM. A Review of the Literature.” *Journal of Technology Management & Innovation* 14: 97–106.
Koster, Ferry and Luc Benda. 2020. “Innovative Human Resource Management. Measurement, Determinants and Outcomes.” *International Journal of Innovation Science* 12, forthcoming.
Koster, Ferry, Marthe Korte, Petra van de Goorbergh and Daan Bloem. 2017. *Innovative HRM Survey. Descriptive Results*. ICOON Paper 10. Tilburg: ICOON.
Lam, Alice. 1997. “Embedded Firms, Embedded Knowledge: Problems of Collaboration and Knowledge Transfer in Global Cooperative Ventures.” *Organization Studies* 18: 973–996.
Maine, Elicia, Sarah Lubik and Elizabeth Garnsey. 2012. “Process-Based vs. Product-Based Innovation: Value Creation Nanotech Ventures.” *Technovation* 32: 179–179.
Nooteboom, Bart. 1994. “Innovation and Diffusion in Small Firms: Theory and Evidence.” *Small Business Economics* 6: 327–347.
Nohria, Nitin and Ranjay Gulati. 1996. “Is Slack Good or Bad for Innovation?” *Academy of Management Journal* 39: 1245–1264.
Oerlemans, Leon A.G., Marius T.H. Meeus and Frans W.M. Boekema. 1998. “Do Networks Matter for Innovation? The Usefulness of the Economic Network Approach in Analysing Innovation.” *Tijdschrift voor Economische en Sociale Geografie* 89: 298–309.
Pouwels, Ivan and Ferry Koster. 2017. “Inter-Organizational Cooperation and Organizational Innovativeness. A Comparative Study.” *International Journal of Innovation Science 9*(2), 184–204.
Pittaway, Luke, Maxine Robertson, Kamal Munir, David Denyer and Andy Neely. 2004. “Networking and Innovation: A Systematic Review of the Evidence.” *International Journal of Management Reviews* 5–6: 137–168.
Rooks, Gerrit, Werner Raub, Robert Selten and Frits Tazelaar. 2000. “How Inter-Firm Co-operation Depends on Social Embeddedness: A Vignette Study.” *Acta Sociologica* 43: 123–137.
Scott, W. Richard and Gerald F. Davis. 2007. *Organizations and Organizing: Rational, Natural and Open Systems Perspectives*. New York: Pearson Education Inc.
Schilling, Melissa A. and Corey C. Phelps. 2007. “Interfirm Collaboration Networks: The Impact of Large-Scale Network Structure on Firm Innovation.” *Management Science* 53: 1113–1126.
Tansky, Judith W. and Robert Heneman. 2003. “Guest Editor’s Note: Introduction to the Special Issue on Human Resource Management in SMEs: A Call for More Research.” *Human Resource Management* 42: 299–302.

Ulrich, Dave and James H. Dulebohn. 2015. "Are We There Yet? What's Next for HR?." *Human Resource Management Review* 25: 188–204.
Van Gils, Anita and Peter Zwart. 2004. "Knowledge Acquisition and Learning in Dutch and Belgian SMEs: The Role of Strategic Alliances." *European Management Journal* 22: 685–692.
Von Krogh, Georg and Nina Geilinger. 2014. "Knowledge creation in the eco-system: Research imperatives." *European Management Journal* 32: 155–163.

Anne Roeters, Esther de Ruijter and Tanja van der Lippe

21 A Transaction Cost Approach to Informal Care

Abstract: Research on cooperation and care has largely overlooked the informal care for adults. Informal care is the care for those who experience (mental or physical) health issues. In this contribution we aim to explain the provision of informal care from a transaction cost approach. We do so by investigating the role of coordination problems and trust problems in the supply of informal care from the perspective of the care giver. We also investigate the role of the social embeddedness of the relationship between the care giver and receiver. Using information from 7,166 care givers and non-care givers collected by the Dutch Institute for Social Research and the Central Bureau of Statistics, multivariate analyses are used to test our hypotheses. Results show that less hours of informal care are provided when the complexity of needs is higher. However, unexpectedly, those with more general skills spend less rather than more time of informal care. Our results also suggest that care givers prefer to give informal care to the ones they know and have a close relationship with. Although the findings are mixed, we conclude that informal care provided by the care giver can be viewed upon as a transaction, and give suggestions for further research.

21.1 Introduction

Cooperation and conflict arise in many different contexts and the family domain is one of these. There is an extensive body of literature that has studied how partners in a household divide and negotiate paid and unpaid work (Blood Jr and Wolfe 1960; Hook and Wolfe 2011; Becker 1981; Becker and Moen 1999; Poortman and Van der Lippe 2009), and how organizations coordinate how parents can spend time on their children (Roeters 2010). The unpaid work activities that are studied are usually limited to household work and child care (Bianchi and Milkie 2010; Bianchi et al. 2012), and studies tend to focus on the relationship between heterosexual partners.

Note: This study is part of the research program Sustainable Cooperation – Roadmaps to Resilient Societies (SCOOP). The publication has benefited from the support of the Netherlands Organization for Scientific Research (NWO) and the Dutch Ministry of Education, Culture and Science (OCW) in the context of its 2017 Gravitation Program (grant number 024.003.025).

Anne Roeters, Netherlands Institute for Social Research
Esther de Ruijter, Arbeid Opleidingen Consult
Tanja van der Lippe, Utrecht University, Utrecht

https://doi.org/10.1515/9783110647495-021

By focusing on these activities and relationships, research on cooperation and care has largely overlooked the informal care for adults. Informal care is the care for those who experience (mental or physical) health issues. This can concern the care for an ageing parent, but also applies to the care for a sibling or neighbor with a disability. The Dutch term for this type of care – 'mantelzorg' translates as 'cloak care', indicating that the person who is taken care of is taken under the wings of the person providing the care. The lack of attention for informal care is a missed opportunity, both from a scientific and a societal point of view.

First, it is a missed opportunity for research studying interdependencies and cooperation (Raub and Weesie 2000). When informal care is provided (or withheld), multiple "negotiations" have taken place, implicitly or explicitly. When a person is in need of care it is not self-evident who provides this care. When household and childcare tasks are involved it is usually evident that both partners share this responsibility. The division of responsibilities is less evident when it comes to informal care. Partner, parents, siblings, neighbors, and friends are all potential informal care givers. In informal care, personal relationships are intertwined with caring tasks. This may provide benefits because there is trust, but it may also be considered a risk because it may have (negative) consequences for the personal relationship. Informal care creates social obligations, and the care receiver is strongly dependent on the care giver (the obligation has to be "paid" to the same person). Thus, there are similarities with outsourcing childcare where the care that is received is also dependent on the care giver (De Ruijter 2005). Moreover, formal care is often an attractive alternative (more so than usually is the case for child care; see Portegijs, Boer, and Merens 2015). A large proportion of the Dutch considers the government to carry the main responsibility for the care of those in need (van den Broek, Dykstra, and van der Veen 2015). The provision of informal care does not only require coordination between those who are receiving and providing this care. When there are multiple care givers (e.g., a neighbor and a child, or a partner and a nurse) they have to coordinate their activities. Naturally, the person who is in need of care has an important say in how this care is arranged, but at the same time he or she is dependent on the availability and willingness of others.

Second, informal care is increasingly important from a societal perspective. Like other western societies, the Dutch population is aging. Currently, 19% of the Dutch population is older than 65 years and the Dutch Bureau of Statistics expects this percentage to increase to 26% in 2040 (Stoeldraijer, Van Duin, and Huisman 2017). But not only older people are in need of care. Physical and mental health impairments can arise at a much earlier age. The average 'healthy life expectancy' (the estimated number of years during with people live in good health) currently varies between 57,2 for lower educated and 71,5 for higher educated (estimates for individuals born in 2017, Centraal Bureau voor de Statistiek (Statline) 2018).Thus it is not surprising that informal care is increasingly common. In the Netherlands, the number of informal care givers is estimated at 5 million people. This equals one third of those aged 16 or older (De Klerk et al. 2017). In the coming decades this percentage is expected to increase

(Van den Broek et al. 2016). The expected increase in informal care is not only driven by demographic changes. Public health policy in the Netherlands is increasingly stimulating informal care. In order to do so, access to formal care is restricted and care professionals and public servants are required to discuss the possibilities for informal care with the family members of those who are in need of care (Broek 2013).

Although there is an increasing body of literature on informal care for adults with health issues, the effects of coordination and trust problems have not yet been addressed. We argue that trust plays a key role in the supply of informal care. Our hypotheses are informed by two theoretical approaches: the *transaction cost approach* (Coase 1952; Williamson 1981, 1985) and *new economic sociology* (Granovetter 1985; Smelser and Swedberg 1994). Both approaches can help us to assess the influence of trust problems on informal care giving. The transaction cost approach describes the influence of trust problems or "opportunism problems" on decision-making by firms. The transaction cost approach has been applied to the family before, usually in combination with insights from new home economics. Although this research has not yet focused on informal care it does focus on issues concerning contracting and financial arrangements in intimate relationships or the outsourcing of household and caring tasks (Ben-Porath 1980; De Ruijter, Van der Lippe, and Raub 2003; Giesen 1999; Ludwig-Mayerhofer 2000; Pollak 1985; Treas 1991, 1993; Treas and Widmer 2000). These studies suggest that the exchange of support can be hindered by trust issues. New economic sociology complements this approach by arguing that trust inspired by social embeddedness reduces risks associated with the exchange of support.

The current study aims to explain the provision of informal care from a transaction cost approach. We do so by investigating the role of coordination problems and trust problems in the supply of informal care from the perspective of the care giver. We also investigate the role of the social embeddedness of the relationship between the care giver and receiver. The dataset that is used to test our hypotheses provides information about the characteristics of the potential care giver (general characteristics as well as characteristics that relate to the ability and willingness to provide care), the care receiver and the relationship between the two. Therefore the data enable us to study the specifics of the context in which care is provided or withheld.

21.2 Theoretical framework

The basic idea of the transaction cost approach is that that governance structures are chosen in such a way that the anticipated costs for reaching and enforcing agreements during transactions are minimized (Coase 1952; Williamson 1981, 1985). Firms can protect themselves from problems by choosing a certain governance structure, such as the detailed contractual planning of a transaction, or by looking for a reliable partner, which involves transaction costs. The properties of a transaction determine

which governance structure is the least costly (Coase 1952; Williamson 1981, 1985). If a firm is more likely to encounter problems when entering a transaction on the market and the damage it can suffer is higher, the firm will incur higher transaction costs to prevent problems. Coordination problems within firms may also encourage market exchange and prevent internalization of certain activities (e.g. Baron and Kreps 1999).

Regarding informal care, the supply of care requires investments in transaction costs. The *likelihood* and potential *consequences* of coordination and trust problems both influence the supply of informal care. In the literature, these two elements of trust problems are described as the problem potential of a transaction (Batenburg et al. 2003). The higher the problem potential, the more costs are needed to prevent problems (e.g. low-quality care, negative effects of informal care on the quality of the personal relationship). As a consequence, care givers may refrain from supplying informal care due to the high expected costs associated with the exchange. Therefore, we expect that a higher problem potential has a negative effect on the supply of informal care and the investments made by the care giver in the informal care relation.

Trust problems in informal care relate to the competence, values and opportunism. We focus on the perspective of the care giver. If suppliers feel that they are not competent enough, they may experience feelings of stress because they feel unable to supply the required care. Regarding values, a care giver can perform a task unsatisfactory due to different standards of hygiene or cleanliness of the care giver and care receiver. This may increase costs from the perspective of the care giver. Also, a care receiver may behave opportunistically, for instance by taking advantage of the care giver (e.g. increasingly claiming time). These types of problems may also exist from the perspective of the care receiver. However, we focus in this study on the problem potential experienced by the care giver. The higher the problem potential, the higher the expected costs for the care giver to prevent problems and the less inclined the supplier is to provide the informal care.

Coordination problems arise from difficulties related to combining work, home and care, and the extent to which care givers need to adjust their activities to others (Treas 1993). Transaction costs are incurred to "reduce day-to-day hassles of negotiating and coordinating exchanges (i.e. to avoid distasteful haggling, minimize unpleasant disputes, eliminate awkward misunderstandings, cut down the time wasted policing the performance of others)" (Treas 1993: 724). The transaction costs of informal care are higher when there are bigger coordination problems to deal with. Coordination problems are more likely to arise when multiple roles at work and in the family have to be synchronized (e.g. Voydanoff 1987, 1988). For example, a demanding job with long working hours and little flexibility, may make it difficult to attend doctors appointments or provide other types of support and care. The more coordination problems associated with the informal care, the less likely the care giver is to provide informal care.

The *new economic sociology* addresses the effect of the embeddedness of the relationship between the buyer and supplier in transactions (Granovetter 1985; Smelser and Swedberg 1994). This embeddedness argument emphasizes "the role of concrete

personal relations and structures (or 'networks') of such relations in generating trust and discouraging malfeasance" (Granovetter 1985: 490). Informal care involves risks that can be mitigated by the social embeddedness of transactions. The social embeddedness of the informal care relation induces trust and reduces the required transaction costs. A greater embeddedness of care supply in social relations provides information, for instance about values, skills or expectations regarding the informal care. It also allows for effective non-legal rewards and sanctions, for instance if something goes wrong or if the care receiver takes advantage of the care giver. Therefore, a greater embeddedness reduces the required investments in transaction costs to prevent problems.

The social embeddedness of transactions has a dyadic and a network aspect (Granovetter 1985; Raub and Weesie 1990). The *dyadic embeddedness* of a transaction refers to the ongoing character of a dyadic relationship. *Network* embeddedness is the extent to which actors are linked to third parties in a social network (Raub and Weesie 1990). Both types of embeddedness provide the care giver information ("learning") as well as possibilities for sanctioning ("control") (Buskens 2002, Buskens and Raub 2002).

Based on the literature, we expect that a higher problem potential has a negative effect on the supply of informal care. Therefore, factors that increase the problem potential are expected to reduce the frequency of care activities and amount of time spent providing care. Table 21.1 provides an overview of the specific characteristics of the 'care situation' that are expected to impact the problem potential and provision of care. First, the complexity of the care needed is expected to increase the problem potential and decrease the supply and investments, because there is a higher risk of things going wrong and the consequences for the care giver are more severe. Second, we expect higher skills of the care giver to reduce the problem potential because fewer investments are needed to provide the informal care and the care giver will experience less stress in providing the care. Third, time demands of the care giver are expected to increase coordination problems and therefore have a negative effect on the hours of informal care and investments in the care relation. Fourth, the social embeddedness and closeness of the relationship of the care giver and care receiver are expected to generate trust and decrease the problem potential and therefore increases the supplied hours of informal care and the investments made in the care relation.

Table 21.1: Expected relations between explanatory variables and hours of informal care.

Explanatory variables	Effect on problem potential	Effect on providing care
Complexity of care needs and situation	+	–
Skills of the care giver	–	+
Time demands of the care giver	+	–
Closeness of relationship care giver and care receiver	–	+

Drawing on the transaction cost approach, one would expect that the association between the care situation and the problem potential and provision of care are conditional on governance structures. However, informal care is not a service that can be the subject of a formal contract or arrangement. It always depends on the willingness of both parties to provide and receive care. One could consider alternative governance structures such as informal agreements between multiple care givers or informal agreements between the care giver and care receiver. For example, decisions regarding the care can be made ad hoc, but they can also be the result of extensive negations. And agreements can be implicit or they can be made explicit and written down as a list of tasks and responsibilities. Unfortunately the data do not allow us to study such variations.

21.3 Methods

21.3.1 Data

The analyses were based on the "Informele Zorg" data that were collected by the Dutch Institute for Social Research and the Central Bureau of Statistics. The sample size was 18,882 persons (Janssen 2017; Klerk et al. 2017). The sample was stratified by region. Because data collection was aimed at informal care (including volunteering) a lower threshold was set for the number of 'active carers' (informal carers and volunteers in the care sector). A minimum of 2,800 carers were required to respond.

Field work took place between September and December 2016. Respondents received a letter that invited them to fill out a web-survey. Those who had not responded after two reminder-letters and with an available telephone number, were approached by telephone. When the respondent was contacted and willing to participate, a telephone interview was held. The overall response rate was 38%. A total of 7,166 individuals responded; 2,852 of them qualified as 'active carers'. The data collection takes the perspective of the care giver as starting point, and implies that we can only test our hypotheses for the group of care givers and are not able to compare this with the group who does not provide any care at all (see for more information analytical strategy).

We have extensive information about the care giver and the one who receives care. However, for those who indicated that they do not provide informal care, we have much less information. We have information about their socio-economic background characteristics and attitudes towards care, but we do not know if they have someone in their social network who is in need of care. This implies that we can only test hypotheses on the role of transaction costs for the intensity of care. That is, we can compare informal carers with varying levels of time investments in informal care, but we cannot compare informal carers with those who do not provide any care at all. Because it is possible that those who provide informal care are a

selective group, our analyses briefly investigate how the background characteristics of informal carers differ from those who do not provide informal care.

21.3.2 Measures

Independent variables. The *complexity of care needs* was measured with two indicators. First, we included a dummy-variable indicating whether the care receiver could be left alone for more than 30-minutes (1 = *always*, 0 = *often, sometimes, seldom or never*). Second, we created a count-variable measuring the number of conditions that were relevant for the care receiver's care needs. The respondents were able to select one or multiple conditions from a list of nine conditions (temporary physical disability, chronic physical disability, terminal illness, dementia and related diseases, mental disorder, psycho-social problems, mental disabilities and others). Assuming that providing care is more complex when multiple conditions co-occur, we counted the total number of conditions.

Three variables measure the *level of knowledge and skills* of the care giver. First, the questionnaire included the question "Have you ever worked in health care or social work and given assistance or help to clients or patients?". We assume that those with experience in this sector will find it easier to provide care because they have been trained to do so. The second and third variable measure the self-perceived skills and knowledge. Respondents were asked four specific and three more general questions. The four specific items are "I remain calm when I encounter difficulties while providing care", "I generally handle unexpected events during care giving well", "If things do not work out when I'm providing help, I find ways to do what is necessary", and "I know where to turn to when I have questions or experience problems with regard to care giving", with answer categories on a Likert scale. The three general questions are: "Do you consider yourself capable, to help the person you provide care for?", "Do you believe you lack knowledge to help the person you provide care for?", and "Do you believe you have the necessary skills to help the person you provide care for?" The yes and no-answers to these items were combined into two separate scales (taking the mean score). The items were coded in such a way that higher scores reflect more skills and knowledge.

The *time demands* on the care giver are assumed to be higher if he or she is in paid employment (0 = *no*; 1 = *yes*), works longer hours (an interval variable measuring the number of hours per week excluding overtime), and lives with a dependent child (0 = *no*; 1 = *yes*). Moreover, we measure the level of subjective time pressure. Respondents are asked "Can you indicate – on a scale from 1 to 10 – to what extent you feel like you are under time pressure in your daily life? A 1 means that you experience 'very little time pressure' and a 10 that you experience 'a lot of time pressure'."

Our indicator of *social embeddedness* distinguishes between two levels of closeness (0 = *distant*; 1 = *close*). The relationship is assumed to be close if the care

receiver is the care giver's partner or close family member (i.e. parent, child, sibling). The relationship is assumed to be more distant for distant family members (e.g. uncle) and non-kin (e.g. neighbor, friend).

Finally, we control for gender and attitudes towards informal care. *Gender* is measured with a simple dummy-variable (0 = *men*; 1 = *women*). The scale measuring *attitudes towards informal care* is based on four statements. "When parents need a lot of help, they should be able to live with their children"; Neighbors have the duty to help other neighbors when they are in need of help; Family members should help other family members when they have health problems or impairments; Friends should help other family members when they have health problems or impairments (answer categories range from 1 = completely agree to 5 = completely disagree). The alpha is 0.74. The items were reverse-coded and combined into one scale. On the new scale (ranging from 1 to 5) higher values reflect more positive attitudes towards providing informal care.

Dependent variables. In the first step, we analyze the difference between the respondents who do and do not provide informal care. Respondents are identified as informal carers if three conditions are satisfied: whether (i) they have provided informal care in the last 12 months, (ii) they currently provide informal care *and* (iii) the main person they provide care for is 18 years or older. Because people do not always recognize that they provide informal care, the question that identifies informal care has an elaborate introduction: "The following questions concern the provision of help to social contacts with health impairments. For example, consider your partner, family, friend or neighbor who needs help because of physical or mental impairments or his or her age. Examples are household tasks, help with personal hygiene, keeping the him or her company, transportation or manual jobs. Help as part of your job or volunteering does not count." Respondents are considered as informal carers if they answer "yes" to this question. After this introduction the respondent is first asked whether he or she had provided this type of help in de preceding 12 months. Because the care for children has a different dynamic and not even 5% provided this type of care we decided to focus on informal care to adults. Thus we excluded 78 respondents who provided informal care to someone under the age of 18, setting their scores to missing. After this selection 2,066 respondents (27.6%) were labeled as carers and the remaining 5,350 (71.4%) respondents were labeled as non-carers.

After investigating the differences between those who do and do not provide informal care, we test our hypotheses predicting *the level of investments in the informal care relationship*. Two outcome measures are considered: The care giver is asked to estimate the frequency of care (rarely, monthly, weekly, daily) and the number of hours per week. On average the care givers spend 7.7 hours per week providing care. Figure 21.1 shows how frequent the care givers provide care. Care on a weekly basis is most common.

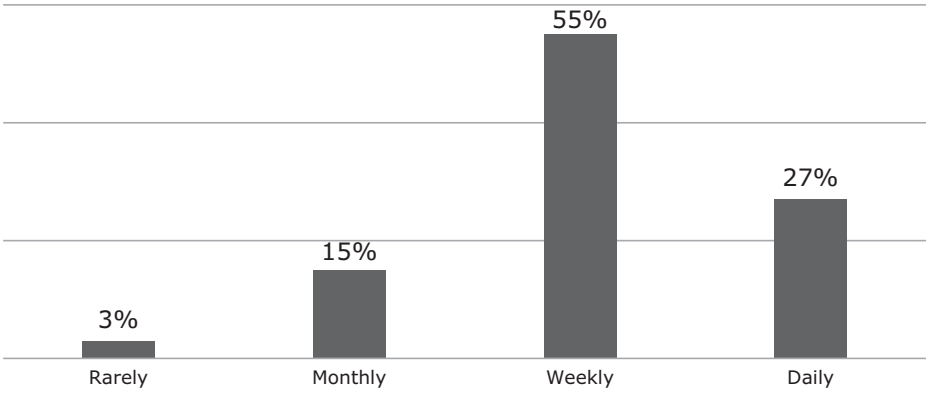


Figure 21.1: Investments in informal care by care giver in frequency.

21.3.3 Analytical strategy

Unfortunately, there is no way of knowing whether those who do not provide care, have a family member or neighbor who is in need of care. Nor do we have information about the relationship, complexity of care needs, and knowledge and skills of the potential carers. This implies that we are largely restricted to studying variations within the group who provides care. Because this is likely to be a selective group, the first step of our analyses compares the background characteristics of the informal carers and the respondents who did not provide informal care.

In the second step we explain the intensity of care. We do so by investigating the role of the complexity of care, the level of skills and knowledge of the care giver, and the social embeddedness. We apply each of these explanations to each of the four outcome measures. Because the outcome measures have different scales, we use different analytical models: (i) The model explaining the frequency of care are ordered logit regression models; (ii) the model explaining the number of hours of care is an OLS-model.

21.4 Results

21.4.1 A comparison between background characteristics of care givers and non-care givers

Are care givers a selective group? Table 21.2 shows that this is not the case with respect to background characteristics such as the labor market situation and attitudes towards informal care. However, it also becomes clear that care givers are more likely to have work experience in the care sector, that they are more likely to have a

Table 21.2: Descriptive results and the comparison of background characteristics for care givers and non-care givers.

	R. does not provide informal care	R. provides informal care	p-value[a]
Complexity of care needs and situation			
Number of conditions		1.3	
Care receiver can always be left alone for >30mins (ref = the care receiver cannot)		69.4%	
Knowledge and skills of the care giver			
Work experience in care	21.1%	32.2%	.000
Scale specific skills and knowledge		4.0	
Scale general skills and knowledge		2.1	
Time demands of the care giver			
R is employed	70.9%	70.9%	.923
Working hours	23.9	23.0	.342
Care giver lives with a dependent child	58.1%	50.0%	.000
Time pressure	5.5	5.9	.000
Individual characteristics			
Female	48.9%	55.6%	.005
Positive attitudes towards informal care (1–5)	3.2	3.1	.558

Note: [a] Estimated on the basis of a logistic regression model explaining the likelihood that the respondent provides informal care (controlling for the other background characteristics).

dependent child and to be female. Moreover, they report higher levels of time pressure than those who do not provide informal care. These descriptive findings are in contrast with the transaction cost theory, where we would expect that labor market situation and attitudes matter for informal care. We should be careful drawing conclusions based on these descriptive results, because we do not take into account the amount of caregiving in Table 21.2. People may in fact be inclined to provide limited care due to social expectations even when a large problem potential is involved, while reducing the amount of care depending on the problem potential. This may reduce the differences when we only compare caregivers and non-caregivers.

21.4.2 Explanatory analyses

Table 21.3 provides the results for the Ordinary Least Squares Regression analysis to explain hours of informal care and Ordinal Logistic Regression to explain the frequency of care provided. We start with the hours of informal care. The results show that, as expected, less hours of informal care are provided when the complexity of care needs is higher: the number of conditions relates negatively to hours of informal care.

Table 21.3: Explanatory analyses to explain the estimated hours[a] and frequency[b] of informal care, informal carers (unstandardized coefficients, weighted analyses).

	Hours	Frequency
Complexity of care needs and situation		
Number of conditions	−2.08**	−0.02
Care receiver can always be left alone for >30mins (ref = the care receiver cannot)	−3.36*	−0.45***
Knowledge and skills of the care giver		
Work experience in care (ref = no experience)	0.41	0.07
Scale specific skills and knowledge	2.37	0.26*
Scale general skills and knowledge	−2.24**	0.02
Time demands of the care giver		
R is employed (ref = nonemployed)	−6.94**	−0.60**
Working hours	0.03	−0.003
Care giver lives with a dependent child (ref. = no child)	−1.06	0.21
Time pressure	−0.3	0.02
Social embeddedness		
Care receiver is partner or close family (ref = distant)	3.83*	1.28***
R is female (ref = male)	1.5	0.07
Positive attitudes towards informal care (1–5)	−0.41	−0.07
Constant	17.57**	
Constant cut 1		−0.56
Constant cut 2		2.32***
R-squared	0.05	–
Number of cases	1,430	1,430

Notes: ** $p < 0.01$; * $p < 0.05$
[a]OLS regression.
[b]Ordered logistic regression.

However, the results also show that when the care receiver can be left alone, less care is provided. Furthermore, if the care giver perceives that he or she has more general skills, unexpectedly less time is spent on informal care. This seems to indicate that skills are less important for hours spent on informal care. Perhaps this could be due to efficiency benefits: skilled caregivers may need less time to provide the care than non-skilled caregivers. Another explanation could be that non-skilled caregivers overestimate their general caring skills. Perhaps differences arise only when focusing on specialized rather than general skills. When the care giver is employed, an indicator of time demands, the results show that less time is spent on providing informal care. Social embeddedness matters with respect to the closeness of the relationship: if the care receiver is partner or close family, more time is spent on informal care.

If we then turn to the frequency of care, it becomes clear that the number of conditions does not matter. Moreover, if the care receiver can be left alone, this leads again to less informal care. Furthermore, for frequency skills do matter, more specific skills imply more care, which is as we would expect. Also time demands,

indicated by employment of the care giver is negatively associated with frequency of care. This result was also found for hours of informal care. Social embeddedness matters as well: if the care receiver is a partner or close family, as expected, more informal care is provided.

21.5 Conclusion

This study investigated the relevance of a transaction costs approach to informal care. Using insights from this approach together with social embeddedness arguments, we were able to formulate hypotheses on hours and frequency of informal care. Capitalizing on rich data on care givers in the Netherlands, we were able to test our hypotheses.

Our main conclusion is that informal care provided by the care giver can be viewed upon as a transaction. When there are more coordination problems in organizing and providing informal care, the problem potential is higher, and this will have a negative relation with the supply of informal care. We therefore fully embrace the idea of informal care as a transaction, because it gives more insight in the decision to provide informal care. Also, our empirical findings partly confirm the idea of informal care as a transaction, at least in the way we have hypothesized about the transaction costs involved in informal care.

Our findings are in line with the idea that when the complexity of care is higher, and when there are more time demands, coordination is more difficult and less informal care will be provided. However, when the care receiver cannot be left alone, typically an indicator of complex needs, more care is provided as well. This result might not be so surprising, since these people are more in need of informal care, but according to the problem potential argument, we expected less informal care. For skills, results are partly as expected. Specific skills matter for frequency of informal care. Nevertheless, those with more general skills spend less rather than more time of informal care. Possible explanations could be that skilled caregivers experience efficiency benefits, non-skilled caregivers overestimate their skills or that differences only become apparent when looking at specialized caring skills.

Our results also suggest social embeddedness is relevant in understanding the transaction between the care giver and the care receiver. When the care receiver is partner or close family, more time is spent on informal care and the frequency is higher as well. Thus, social embeddedness makes a difference in informal care exchanges as well as intra-household exchanges (De Ruijter 2005). Care givers prefer to give informal care to the ones they know and have a close relationship with. The dyadic relation they have inspires trust, and the risks associated with informal care are mitigated by the social embeddedness of the transaction.

All in all, our study provides a promising avenue for further research. We have three suggestions thereby. In this study we were not able to focus on the content of care provided, but we suggest other researchers to delve into this topic, because it might be helpful to explain some of the unexpected findings. It probably matters whether care receivers need complex help (e.g. wound care, administering medicine), involving a high problem potential, compared to simple caring tasks such as help with personal hygiene. In more complex caring situations we would expect larger effects of the problem potential on caregiving, because the consequences of problems are more severe. Furthermore, in this chapter we focused on temporal measures of investments (the frequency of care and hours of care) but care givers can invest in the care relationship also in other ways. Therefore, we advise in new research to take investments in the informal care relationship into account, such as moving house in order to be closer to the care receiver, to make arrangements with the employer or to take up responsibilities that are time-extensive, such as helping with filing taxes once per year. Finally, although we used unique data, it would even be better to have longitudinal data to study the process of informal care, and the transactions that are made over a longer time span. This would enable us to understand the causality of the relation between trust and coordination problems and caregiving. It would also help us to understand how skills can develop over time and thereby may mitigate trust problems. For example, when caregivers start with simple caregiving tasks they may become more competent and confident over time, thereby reducing the problem potential when more complex caring demands arise. It could be that caregivers are more inclined to give more care when the care demand grows gradually compared to a sudden complex care demand (e.g. as a consequence of an accident or an acute serious illness).

References

Baron, James N., and David M. Kreps. 1999. “Consistent Human Resource Practices.” *California Management Review 41*(3):29–53.

Batenburg, Ronald S., Werner Raub, and Chris Snijders. 2003. “Contacts and Contracts: Dyadic Embeddedness and the Contractual Behavior of Firms.” Pp. 135–188 in *Research in the Sociology of Organizations. Volume 20. The Governance of Relations in Markets and Organizations*, eds V. Buskens, W. Raub and C. Snijders. Oxford: JAI/Elsevier.

Becker, Gary. 1981. *A Treatise on The Family. Volume 30*. Cambridge, MA: Harvard University Press.

Becker, Penny Edgell, and Phyllis Moen. 1999. “Scaling Back: Dual-earner Couples’ Work-family Strategies.” *Journal of Marriage and the Family* 61 (4):995–1007.

Ben-Porath, Yoram. 1980. “The F-Connection: Families, Friends, and Firms and the Organization of Exchange.” *Population and Development Review* 6:1–30.

Bianchi, Suzanne M., and Melissa A. Milkie. 2010. “Work and Family Research in The First Decade of The 21st Century.” *Journal of Marriage and Family* 72 (3):705–725.

Bianchi, Suzanne M., Melissa A. Milkie, Liana C. Sayer, and John P. Robinson. 2012. “Housework: Who Did, Does or Will Do It, and How Much Does It Matter?” *Social Forces* 91 (1):55–63.

Blood Jr, Robert O, and Donald M Wolfe. 1960. *Husbands and Wives: The Dynamics of Family Living*. Oxford, England: Free Press.

Boer, Alice de, Mirjam de Klerk, and Ans Merens. 2015. "Mij een Zorg?! Zorgvisies van de Overheid en Burgers Vergeleken." *Tijdschrift voor Arbeidsvraagstukken* 31 (4):494–509.

Broek, Andries van den, Cretien van Campen, Jos de Haan, Anne Roeters, Monique Turkenburg, and Lotte Vermeij. 2016. *De Toekomst Tegemoet*. Den Haag: Sociaal en Cultureel Planbureau.

Broek, Thijs van den. 2013. "Formalization of Informal Care in The Netherlands: Cost Containment or Gendered Cost Redistribution?" *IJFAB: International Journal of Feminist Approaches to Bioethics* 6 (2):185–193.

Broek, Thijs van den, Pearl A. Dykstra, and Romke J. van der Veen. 2015. "Zorgidealen in Nederland: Verschuivingen tussen 2002 en 2011." *Mens en Maatschappij* 90 (1):25–52.

Buskens, Vincent. 2002. *Social Networks and Trust.* Boston/Dordrecht/London: Kluwer Academic Publishers.

Buskens, Vincent, and Werner Raub. 2002. "Embedded Trust: Control and Learning." *Advances in Group Processes* 19: 167–202.

Centraal Bureau voor de Statistiek (Statline). 2018. *Gezonde Levensverwachting: Onderwijsniveau*. Retreived from: https://opendata.cbs.nl/statline/portal.html?_la=nl&_catalog=CBS&tableId=83780NED

Coase, Ronald H. 1952. "The Nature of the Firm." Reprinted in *Readings in Price Theory*, eds. G. J. Stigler and K.E. Boulding. Homewood, IL: Richard D. Irwin.

Giesen, Deirdre. 1999. "Juridische Arrangementen." Pp. 55–80 *in Huwelijks- en Samenwoonrelaties in Nederland. De Organisatie van Afhankelijkheid*, eds. Matthijs. Kalmijn, Wim Bernasco and Jeroen Weesie. Assen: Van Gorcum.

Granovetter, Mark. 1985. "Economic Action and Social Structure: The Problem of Embeddedness." *American Journal of Sociology* 91:481–510.

Hook, Jennifer L, and Christina M Wolfe. 2011. "Parental Involvement and Work Schedules: Time with Children in the United States, Germany, Norway and the United Kingdom." *European Sociological Review* 29 (3):411–425.

Janssen, Björn. 2017. *Onderzoeksverantwoording Dataverzameling Informele Zorg 2016*. Den Haag: Centraal Bureau voor de Statistiek.

Klerk, Mirjam de, Alice de Boer, Inger Plaisier, and Peggy Schyns. 2017. *Voor Elkaar? Stand van Informele Hulp in 2016*. Den Haag: Sociaal en Cultureel Planbureau.

Ludwig-Mayerhofer, Wolfgang. 2000. "Transaction Costs, Power, and Gender Attitudes in Financial Arrangements of Couples." Pp. 46–47 in *The Management of Durable Relations: Theoretical Models and Empirical Studies of Households and Organizations*, eds Jeroen Weesie and Werner Raub. Amsterdam: Thela Thesis.

Pollak, Robert A. 1985. "A Transaction Cost Approach to Families and Households." *Journal of Economic Literature* 23:581–608.

Poortman, Anne-Rigt, and Tanja van der Lippe. 2009. "Attitudes Toward Housework and Child Care and The Gendered Division of Labor." *Journal of Marriage and Family* 71 (3):526–541.

Portegijs, Wil, Alice de Boer, and Ans Merens. 2015. Mij een zorg!? *Tijdschrift voor Arbeidsvraagstukken* 31(4): 494–509.

Raub, Werner and Jeroen Weesie. 1990. Reputation and efficiency in social interactions: An example of network effects. *American Journal of Sociology* 96: 626–654.

Raub, Werner, and Jeroen Weesie. 2000. "The Management of Matches: A Research Program on Solidarity in Durable Social Relations." *The Netherlands Journal of Social Sciences*, 36(1): 71–88.

Roeters, Anne. 2010. *Family Life Under Pressure? Parents' Paid Work and The Quantity and Quality of Parent-Child and Family Time*. PhD thesis, Utrecht University.

Ruijter, Esther de. 2005. *Household Outsourcing*. PhD thesis, Utrecht University.
Ruijter, Esther de, Tanja van der Lippe, and Werner Raub. 2003. "Trust Problems in Household Outsourcing." *Rationality and Society* 15(4):473–507.
Smelser, Neil J., and Richard Swedberg. 1994. *The Handbook of Economic Sociology*. Princeton, NJ: Princeton University Press.
Stoeldraijer, Lenny, Coen van Duin, and Coen Huisman. 2017. "Bevolkingsprognose 2017–2060: 18,4 Miljoen Inwoners in 2060". In *Statistische Trends*. Den Haag: Centraal Bureau voor de Statistiek.
Treas, Judith. 1991. "The Common Pot or Separate Purses: A Transaction Cost Interpretation." Pp. 211–224 in *Gender, Economy, and Family: The Triple Intersection*, ed. R. L. Blumberg Newbury Park, CA: Sage.
Treas, Judith. 1993. "Money in the Bank: Transaction Costs and the Economic Organization of Marriage." *American Sociological Review* 58: 723–734.
Treas, Judith, and Eric D. Widmer. 2000. "Whose Money? Financial Management in Marriage: A Multi-Level Analysis for 23 Countries." Pp. 44–45 and CD In *The Management of Durable Relations. Theoretical Models and Empirical Studies of Households and Organizations*, eds Jeroen Weesie and Werner Raub. Amsterdam: Thela Thesis.
Voydanoff, Patricia. 1988. "Work role characteristics, family structure demands, and work/family conflict." *Journal of Marriage and the Family* 50 (3):749–761.
Voydanoff, Patricia. 1987. *Work and Family Life*. Thousand Oaks: Sage Publications, Inc.
Williamson, Oliver E. 1981. "The Economics of Organization: The Transaction Cost Approach." *American Journal of Sociology* 87:548–77.
Williamson, Oliver E. 1985. *The Economic Institutions of Capitalism*. New York: The Free Press.

Beate Volker

22 Trust is Good – Or is Control Better? Trust and Informal Control in Dutch Neighborhoods – Their Association and Consequences

Abstract: The idea of collective efficacy – the degree to which residents engage in collective good production and protection – has been established as key for the understanding why neighborhoods sometimes fail in establishing social and physical order. Theoretically, a basic assumption is that collective efficacy rests in the close association between informal social control and trust. This paper argues that this alleged link between control and trust is not always present and even not always plausible. Different possible relationships between trust and control are discussed and empirically explored by multilevel models of behavior in neighborhoods, and it is examined to which degree control and trust go with important neighborhood consequences such as networks among neighbors, collective action with neighbors, and general satisfaction with the neighborhood. Data from the SSND (Survey of the Social Networks of the Dutch 2014, n = 1067, in 165 neighborhoods) are used. Findings show that trust and informal control are only modestly associated with each other. Furthermore, effects of control are not robust in the statistical models, while trust effects are. Finally, control and trust alignment in different neighborhoods is explored and it is argued that the wider neighborhood context such as type of houses, degree of urbanization and neighborhood history influence the degree to which control goes together with trust.

22.1 Introduction: Collective efficacy and trust-control alignment

Trust is good, but control is better.[1]

In their seminal paper on the explanation of neighborhood crime and disorder Sampson, Raudenbush, and Earls (1997) argued that collective efficacy is the key for the understanding of how neighbors safeguard collective good production such as

1 This phrase is ascribed to Vladimir Lenin, although there is no source proving that it has been literally forwarded by him. It comes close to a popular Russian saying "Dowjerjaj, no prowjerjaj" – which means 'be trusting, but verify'. Probably, Lenin used that saying too.

Beate Volker, Department of Human Geography and Spatial Planning, Utrecht University

https://doi.org/10.1515/9783110647495-022

safety, cleanness, and livability of a neighborhood in general. They demonstrated empirically that in neighborhoods where collective efficacy is high, incidents such as registered violence and rates of homicides as well as ratings of perceived violence are lower than in neighborhoods with low collective efficacy. Ever since then, the idea of a lack of collective efficacy – the shared norm that collective goods are protected by interventions of local residents – is used to explain the functioning (or the malfunctioning) of neighborhoods.

The core parts of 'collective efficacy' are social cohesion and perceptions of trustworthiness together with informal control. Collective efficacy is conceived as a composite of trust and informal control, and both are seen as depending on each other. For example, Sampson, Morenoff, and Earls (1999: 919) argue that high trust and cohesion in neighborhoods provides the most "fertile contexts for the realization of informal control". The measurement of collective efficacy is usually straightforwardly based on indicators of trust or cohesion[2] and combined in one scale, together with items measuring informal control. In most studies on neighborhoods, informal control and trust perceptions are considered to be closely intertwined.

However, the extent to which trust and (informal) control are associated with each other is also debated, in particular in the organizational literature (see for example Bijlsma-Frankema and Costa, 2005; Vlaar, Van den Bosch and Volberda 2007). If trust is understood as the expectation that 'actions of others will be beneficial rather than detrimental' (Gambetta 1988) and control as actively monitoring behavior of others (Janowitz 1991), control and monitoring are not necessary in trustful relationships. On the contrary, in a situation where many trustful relations are present, explicit control can be interpreted as a signal of low trust.[3] In other words, if control is carried out actively, it can actually be a substitute of trust. In cases where trust is high, control[4] is not necessary. The opposite also holds: when trust is low, collective good production cannot be warranted without control. Situations where control substitutes trust are quite common and part of rationalization in society. Think for instance of the detailed registrations of employee activities in occupational sectors such as health or education. In fact, control is also an important part of many state

2 In this contribution, the difference between 'trust' and 'social cohesion' is not considered relevant. 'Trust' as used in this chapter refers to dyadic relationships, while 'social cohesion' is a characteristic of a group. However, highly cohesive networks usually consist of trustful and trustworthy relationships and, perhaps more important, most scales on neighborhood cohesion measure trust among neighbors and aggregate the average scores to the level of the neighborhood.

3 I am aware that the situation is more complex, though. On a second level of the interaction, control does play a role. For example, it has to be controlled whether the behavior was indeed in conformation with the agreed norms. Still, however, the argument holds that trust and control cannot be seen as closely related by default.

4 For reasons of brevity I refer to 'control' although my empirical focus as well as my arguments are on 'informal' control. Formal rules and mechanisms are not considered here, since my empirical case, that of neighborhoods lacks these regulations.

systems as in politics, incentives to abuse power are often large and the consequences of such an abuse undesirable. As Warren (1999) puts it: "An important democratic innovation was the recognition that in many relationships trust is misplaced or inappropriate, suppressing real conflicts of interest" (Warren 1999: 1). In short, the abovementioned Russian saying (footnote 1) has been taken seriously in modern western societies.

However, trust and control can also be related in another way. Control mechanisms can increase trust because the rules of the situation are, with the control mechanisms in place, clear as well as the assessment and evaluation formats. If control is understood as the possibility to sanction (and not as the monitoring of the sanction itself), trust and control are no antagonists anymore. From a game-theoretical perspective, if there is the possibility to sanction (control), in equilibrium, there is no need to carry out the sanction. Hence, in such a situation, control promotes trust (see, for example Coleman 1990, Buskens and Raub 2002). In addition, also for the reversed relationship, that trust promotes control, arguments are provided (see above, Sampson, Moreno and Earls 1999).

Last but not least, it can also be argued that control and trust are two conditions in social situations that operate independently of each other, they might simply be not related in one and the same interaction situation.[5] Whether control or trust is important for neighborhood social order might depend on the neighborhood matter in question, the organization of a barbeque might be only related to trust in the neighbor networks, while the arrangement on noise, littering, and car parking might be subject to control.

Given these considerations it is actually puzzling that the literature on collective efficacy in neighborhoods does not at all problematize the relationship between control and trust. Instead, it is usually implied that they are equally important and at play at the very same moment. The established conceptualization and the measurement of collective efficacy assume a coincidence of both, trust and control. Hence, while the link between neighborhood disorganization and neighborhood crime is understood in particular through the mediating effect of collective efficacy (Sampson et al. 1999; Sampson 2006), it remains unclear how the constituents of collective efficacy are related to each other. They might be each other's substitute, supplement (and even reinforce each other's effect), or be unrelated.

Knowledge about the relationship between the elements of collective efficacy is important if we want to understand how neighborhoods function and what conditions promote social order. Do people sense and initiate monitoring or do they just trust, without further 'back-up'? Furthermore, does this equally hold for different

5 The situation might be even more complex, though. It might be that in a given situation an initial amount of trust determines the amount of control exercised.

aspects of neighborhood social order, for instance for actual relationships as well as for actions towards collective good production?

This paper aims to contribute to the disentanglement of the relationship between (informal) control and trust, while studying their consequences for a number of different outcomes: actual neighborhood relations, collective action in neighborhoods, and the general satisfaction with the neighborhood. I focus on the alleged constituents of collective efficacy – trust and control – examine how they are related with each other, and how they contribute to different aspects of neighborhood functioning. In addition, I explore some cases where control and trust are both high, where they are both low, as well as where they do not align. Hence, the research question of this paper is '*what is the association between informal control and trust in neighborhoods, and to what degree do they explain neighbor networks, collective activities, and neighborhood satisfaction?*' Data from the 3rd wave of the SSND (Social Survey of the Networks of the Dutch) are used in combination with key-figures of neighborhoods provided by Statistics Netherlands.

22.2 Collective efficacy, trust, and control – arguments and expectations

22.2.1 Collective efficacy

The idea of collective efficacy and its consequences for a variety of socially desirable outcomes is based on Bandura's psychological theory of individual-level 'personal efficacy' or 'self'-efficacy. This type of efficacy is the belief of individuals that they can attain goals through their own actions – in other words, one's own actions are considered effective for goal attainment (Bandura 1997/2000). Bandura acknowledged that there is also an efficacy belief on a group level: the shared belief in collective power to attain a desired goal. The idea of collective efficacy has been extensively elaborated and applied in sports competitions, but also in educational research (Goddard et al. 2004; Goddard 2001).

In sociology, Sampson, Raudenbush, and Earls (1997) applied the idea of collective efficacy to the study of (dis)functioning neighborhoods, while building upon social disorganization theory. Since the 1970s, social disorganization has been perceived as a community's inability to realize common values and maintain social control (Sampson and Groves 1989; Shaw and McKay 1942). The theory also pointed at three aspects of neighborhood social composition that are expected to enhance disorganization: (ethnic) heterogeneity, residential fluctuation, and disadvantages such as poverty, but also broken families. Consequently, many studies on neighborhood functioning included these three structural neighborhood characteristics in their theoretical and empirical analysis. However, social disorganization theory was

criticized for being a macro level theory and not connected to human actions (Bursik 1988). What people do in order to establish a well-functioning neighborhood remained an open spot in the theory. Sampson's studies (starting with the abovementioned study by Sampson, Raudenbush and Earls 1997) build upon that criticism of social disorganization theory and argued that the mechanism through which the three compositional characteristics actually work can be found within people's relationships and neighborhood networks. Trust as well as informal control are located in such social networks and they constitute the very base for neighborhood social order. Indeed, both, trust and the shared belief that neighbors will intervene on behalf of the common good – that is, collective efficacy – are associated with low rates of crime, ranging from burglaries to violent offenses and even murder. Sampson et al. (1997) showed that collective efficacy substantially mediates the relationship of cumulated disadvantages, fluctuation, and ethnic heterogeneity in neighborhoods, with neighborhood crime rates and violence acts. Hence, collective efficacy hampers the occurrence of collective bads.[6]

The notion of collective efficacy also met with criticism. It has been questioned whether trust and cohesion are really the constituents of collective efficacy in all circumstances (cf. St. Jean 2007, see also Warner and Rountree 1997). The shared belief that people will intervene on behalf of the common good might be sufficient to enhance neighborhood functioning and the degree to which this shared belief requires trust and cohesive networks can be debated.

In addition, arguments have been brought forward concerning the relationship between trust and control: do they really belong to the same phenomenon? For example, Bursik (1999) argues that social control is an outcome of networks – dense networks produce social control as a collective good and, consequently, in such neighborhoods crime rates are low. However, and as mentioned above, if control is the consequence of trust (or vice versa), research should take into account the chronology in the emergence of trust and control, rather than putting them together in one scale and assume that they are present at the same moment. Last but not least, some researchers showed empirically that the items related to trustful relationships among neighbors did not improve the scale of collective efficacy (see Gau 2014; Volker et al. 2015).

In short, there seem to be inconsistencies in the theoretical arguments on which collective efficacy is based upon as well as in the findings acquired from empirical studies. First, the idea of collective efficacy is explicitly based on trust and control – allegedly at play in densely connected, trustful networks. These are networks of stronger ties – and not of weaker ties, given that weaker ties are perceived as less trustful.

6 Volker et al. (2015) showed that collective efficacy also mattered for collective goods: in their study on lost letters in Dutch neighborhoods, letters found in neighborhoods with high collective efficacy were more often returned than the letters found in neighborhoods with low collective efficacy.

However, it has been shown that strong ties can also hamper control efforts (Bellair and Browning, 2010; Browning, Dietz and Feinberg 2004; Flache 1996) because friends – strong relationships – do not easily take action in case of one's criminal or antisocial behavior. They do not intend to put their relationships at risk. Likewise, it has been shown that dense networks of relationships can hamper collective good production rather than enforce it (Wilson 1996; Morenoff, Sampson and Raudenbush 2001). Second, it has been shown repeatedly that network relationships in a neighborhood are – on average – weak rather than strong (see Volker, Flap and Lindenberg 2006; Marsden 1987), at least when one considers network relationships as the degree of closeness among the interaction partners. Neighbors are not the relationships people talk with about their personal matters and people do not ask their neighbors for advice in important decisions or when they have problems in their relationships. Hence, densely connected networks can have opposite effects than the argument suggests, and the argument does not hold for weakly connected networks – such as the networks commonly found in neighborhoods that are nevertheless functioning quite well. In other words, collective efficacy assumptions might not apply in neighborhoods, nor might they apply in networks of strong ties in general. If that is true, trust and control cannot be the mechanisms that explain neighborhood social order.

22.2.2 Trust and control

Although taken for granted in the literature on collective efficacy, the relationship between trust and control is intensely discussed in various fields of the social sciences. In particular the literature on organizational studies and rational choice theory has focused on the connection between trust and control (see Bijlsma-Frankema and Cost, 2005, for an overview). In organizational studies, the relationship between trust and control is studied in relation to governance questions and of growing organizations, with more lateral and less hierarchical relationships. Here, trust and control are often seen as possible antagonists (Deepak and Murnighan 2002) and many authors argue that trust and control are negatively related, (Handy 1993; Ring and Van de Ven 1994; Inkpen and Curral 1997; Cummings and Bromiley 1996; Guseca and Rona-Tas 2001). In that literature, trust is considered as related to risk taking, the risk that one takes while knowing that there is no control that safeguards cooperative behavior. One expects, but cannot be sure, that the actions of other persons are beneficial (Gambetta 1988). Trust and control are considered as substitutes or even negatively related: if one exists, the other is not necessary any more to produce a collective good, or even stronger, one undermines the effect of the other.

While trust can hardly be understood otherwise than as a state of mind, where one has positive expectations towards the actions of others, understanding informal control is more difficult. On the one hand, informal control is a sign of interest and involvement. Think of the teacher, who reads assignments of students and

comments on them in order to help – while doing this s/he also controls the efforts students have put into the assignment. On the other hand, as already mentioned above, in game theory and rational choice theory, control is understood as the possibility to sanction behavior (cf. Buskens and Raub 2002). In that literature, trust is promoted by control, for instance through sanction possibilities, (but see Mulder et al. 2003). The prominent example is given by Coleman (1988: 98) on the wholesale diamond market, which functions without any written contracts or other guarantees just because of the huge sanction possibilities of the members of the community. Consequently, control (next to learning) is seen as essential for the understanding of trust (see Buskens 1999) and, as argued, control is seen as promoting trust. The idea behind 'control' here takes into account that in many situations there is a short-term incentive to abuse trust, but that this might have undesirable consequences in the future. In situations where the shadow of the future is long and abuse of trust can be sanctioned, individual actions are controlled in the sense that they are remembered and there is a chance of revenge for uncooperative behavior. Reciprocity is conceived as the mechanism underlying this type of control (Blau 1964). Furthermore, the literature on neighborhood and social disorder refers to social control as the "capacity of a group to regulate its members according to desired principles to realize collective [. . .] goals" (Sampson, Raudenbush and Earls, op cit: 918).

Taking together, two perspectives on the relationship between control and trust can be distinguished: the perspective of substitution, and the perspective of supplementation or reinforcement. If one argues from the perspective of substitution, one expects a negative relationship in the sense that high trust does not go together with higher control and vice versa. Even more, trust effects on all kind of outcomes might be weaker in the presence of control. High trust requires only low control and vice versa, in order to secure the collective good (see Dekker 2004; Willamson 1975). Arguing from the perspective of supplementation implies that trust and control are positively related and that the effects of one condition are stronger when the other is present. In other words, trust and control effects reinforce each other. Last but not least, a third perspective can be added: it might depend on the composition of the social setting and on the issue in question how trust and control are related. They both can have independent effects on important outcomes, and the strength of these effects depends on the outcome in question. For example, for feeling safely at night when walking through the neighborhood, it is good to know that neighbors control and monitor what is happening, no matter whether one is trusting anyone. On the other hand, if one has asked his or her neighbors to water the plants during one's holiday, one needs to trust them, whether or not one assumes that they control what is going on in the neighborhood. In summary, I examine three possible relationships between trust and control, that of substitution, of reinforcement and of context dependency.

22.2.3 Does it matter? Consequences of collective efficacy

Collective efficacy has been shown to matter for neighborhood social order. All kinds of incidents such as burglaries or perceived and actual reported violence are lower in neighborhoods where collective efficacy is high (Sampson, Morenoff and Earls 1997, 1999, 2001). In line with the idea of collective efficacy, outcomes for wellbeing and health have also been examined (Sampson 2003; Salanova 2003) – though less often. Mostly, the focus has been on the absence of undesirable outcomes rather than on the presence of desirable ones.

As mentioned, this paper focusses on a number of outcomes that are related to the production of collective goods in neighborhoods. Firstly, and basically, I will consider the number of neighbors in the personal network of residents. Having a vital network in the neighborhood where one lives indicates that one interacts face to face, is present in the neighborhood and possibly involved in all kind of social matters. Secondly, the consequences of trust and control will be examined for collective activities among neighbors. Here I distinguish between (i) social activities such as having coffee together or barbequing in the summer, (ii) activities to keep the neighborhood clean or enhance physical order in the neighborhood, and (iii) activities towards institutions in order maintain neighborhood functioning, such as writing petitions to the municipality etc. Finally, the general satisfaction of residents with their neighborhood is considered as an outcome of collective efficacy – trust and control.

These three outcomes indicate behavioral as well as cognitive aspects of how people engage in their neighborhoods. They can be seen as micro-effects of collective efficacy. Studying these different types of outcomes contributes to the understanding of the differential effects of collective efficacy.

22.2.4 Expectations

Straightforwardly, trust and control can be related as follows: They might correlate positively – as it is usually assumed – and contribute both to the outcome under consideration. Given the arguments provided above, they might even reinforce each other. Furthermore, they can be substitutes for each other, implying a negative interaction or, at least that a high coefficient in one condition requires only a low coefficient in the other one for the same outcome. In such a case, their relationship will be antagonistic. As argued above, they might also have independent effects, depending on the outcome in question and do not interact. The third perspective is that of context dependency of the relation between trust and control.

This paper is explorative and there are no theoretical arguments derived for the existence of one or the other relationship presented here. However, I do expect that the relation between trust and control is not a universal one but varies among

neighborhoods and across the outcomes considered. More precisely contextual and compositional neighborhood conditions are expected to matter for the relation between trust and control. For example, in a neighborhood where most residents work full time, building up trustful relationships might be more important than exercising control in case one wants to initiate collective action. Likewise, in a neighborhood with houses built in a way that residents see each other leaving and entering their homes establishing building trust might deserve special attention while control is almost automatically given because of the design of the built environment. Furthermore, it is plausible to expect that the relationship between trust and control is not the same for the phenomena under study: it can be argued that control is more important for collective action – where freeriding behavior is plausible ('if others do it, why should I?') – and that trust is more important for establishing personal relationships or networks in the neighborhood. Neighborhood satisfaction might depend on both control and trust: in a neighborhood where everyone is satisfied people have trustful relationships and those who are not trusted might be controlled.

22.2.5 A note on reversed causality

Given the dynamic nature of trust and control (Buskens 1999) as well as the dynamic nature of neighborhood activities, it is obvious that collective efficacy stimulates networks as much as it depends on neighbor interaction. Also, collective actions will enhance collective efficacy, in particular if they are successful, and those who are satisfied with their neighborhood in general probably have higher beliefs of collective efficacy. It has been referred to these phenomena as reciprocal feedback (Sampson and Raudenbush 1999: 630). Here, the scope of my argument is not on these reciprocal relationships, but I consider only one side of the loop: neighborhood outcomes depending on collective efficacy, trust and control, respectively.

22.3 Data, measurements, and analytical strategy

Data from the Survey of the Social Network of the Dutch (SSND 2014) were employed for this study. The SSND is a larger research project that started in 1999 and consists of 4 waves of interviews with neighborhood residents. The 2014 wave is the third wave and consists of 1067 respondents of which 578 already participated in 2008 and/or in 1999, and 489 respondents belonged to a refreshment sample in the same neighborhoods plus respondents from a number of newly selected neighborhoods. Neighborhoods were delineated on a 5-position zip code basis, which resulted in 161 relatively small neighborhoods in the sample. Interviewer effects as well as selection in attrition have been examined and indications for neither of

these have been found. The original sample is a neighborhood sample, drawn from the 545 municipalities in the Netherlands, while taking density of the population in the different regions into account. For more information about the sample and the neighborhood delineation, see, for instance, Volker et al. (2015).

22.3.1 Dependent variables

As mentioned above, three kinds of dependent variables have been considered. *The number of neighbors* in the personal network has been established as follows: networks were delineated based on the exchange method (see Fischer 1987). Respondents were confronted with a relatively large number of name-generating questions, such as whom they ask for help with odd jobs in and around the house, with whom they talk about important matters, work with, whom they visit and the like. In total 11 of such questions have been asked. In every question 5 new persons could be mentioned. This way, a list of names is generated and in a second step information about the characteristics of theses alters as well as of their relationships with the respondent (ego) have been gathered. One of these questions was the role relationship between respondent and network member, of which 'neighbor' was one. Furthermore, because we expected that neighbors are not prominent in many daily activities, we asked whether respondents knew their neighbors. In other words, there were two possibilities for neighbors to enter the network, as a reaction of a name generating question and via a straightforward question after direct neighbors. In this study, both 'types' of neighbors are considered as the number of neighbors in the network.

Activities, collective actions with neighbors have been measured as follows: Respondents have been asked whether they have undertaken one of the following activities with their neighbors during the last few months: social activities, activities towards a more livable neighborhood, and activities towards institutions.

For the measurement of *social activities*, respondents were asked whether they had i) coffee, ii) a barbeque or a party together. For the measurement of activities towards a better neighborhood they were asked whether they had undertaken activities that made the neighborhood i) more safe, ii) more clean, or ii) whether they had established any rules or arrangement with each other about parking the cars. Lastly, for the measurement of activities towards institutions – the government or the municipality – respondents were asked whether they had undertaken activities like i) calling the police (together), for example because of adolescents hanging around, ii) writing a petition because facilities were removed of the neighborhood or iii) writing a letter of protest to the municipality.

Satisfaction with the neighborhood was measured with the question 'How satisfied are you in general with your neighborhood?' Respondents answered in a 7-point Likert-scale.

22.3.2 Independent variables

Independent variables are informal control and trust in neighborhoods. These have been measured in a similar way as the original items of collective efficacy (Sampson, Raudenbush and Earls, 1997). Neighborhood trust/cohesion was established with the following items:

- people in this neighborhood have good contact with each other,
- if someone needs help, (s)he can count in the neighbors,
- I would not like to have comparable house in another neighborhood, I like it here,
- if there is something to be done, everyone participates,
- I really belong to this neighborhood,
- if I see someone walking in the street I usually know in which house (s)he lives,
- I trust the people in my neighborhood,
- people in this neighborhood trust each other,
- this is a close neighborhood.

Respondents could agree with these items on a 5-point-Likert scale. The items constitute a scale with a Cronbach's alpha of 0.88.

Informal control is measured as expected interventions, similar as in other scales on collective efficacy. People are asked whether they expect that somebody from the neighborhood would intervene in case

- children hang around skipping lessons
- adolescents spray graffiti
- people in the street have a noisy argument
- they observe an attempt of a burglary
- they observe someone breaking in a car
- children fighting in the street
- the municipality intents to open a center for drug addicted in this neighborhood
- the municipality will take away some benches and a playing ground.

Again, answers could be given in a 5-point-Likertscale. The items constitute a scale with an alpha of 0.89.

22.3.3 Control variables

In all analyses the following individual characteristics were used as controls: sex (women = 0, men = 1), age (on years), country of birth (0= outside of the Netherlands,

1 = Netherlands) and highest education (8 categories). In the analyses of the number of neighbors in the network the total network size is also a control variable.

Furthermore, neighborhood composition with regard to age and ethnic background is a control variable. In addition, it is controlled for poverty, a lack of resources and for residential stability: via the percentage of houses for rent, the percentage of divorces, the percentage of people getting unemployment benefit. Urbanization is also controlled for and measured in categories of population density in a squared kilometer. There are 5 classes of urbanism, more than 2500 addresses per km2 (1), between 1500–2500 addresses (2), 1000–1500 addresses (3), 500–100 (4) and less than 500 addresses per km2 (5). All neighborhood control variables were provided by Statistics Netherlands (Statistics Netherlands 2013/2014).

Table 22.1 provides descriptive statistics of key variables in the analyses. Network size is on average 10.7 persons. It can be seen that people differ considerably in their number of neighbors in the network and that the number in general is not high.

Table 22.1: Descriptive statistics of key variables in the analyses.

	%	mean	sd	min	max	N
Individual level control variables						
Sex (female)	50					1069
Age		52.9	14.3	19	96	1069
Education (1–5)						1068
Only primary school and lower vocational training	25.1					
Only secondary school (Mavo, Havo)	18.2					
Higher secondary school/vocational training	22.2					
Higher vocational training	21.5					
University/postgraduation	13.1					
Migration background (yes)	12.9					1069
Network size		10.7	4.6	1	29	1067
Dependent variables						
N of neighbors in network		2.2	1.42	0	11	1067
Social activities with neighbors (count)		.76	.79	0	2	1069
Activities to enhance the neighborhood (count)		.18	.43	0	4	1069
Political activities (count)		.45	.84	0	2	1069
Neighborhood satisfaction (1–7)		5.77	1.10	1	7	1050
Independent variables						
Cohesion (scale/n of items)		3.69	.61	1	5	994
Informal control expectations (scale/ n of items)		3.93	.69	1	5	917
Neighborhood level control variables						
% above 65 years of age	17%			2%	86%	161
% non-western foreigners	17%			0%	89%	161
% divorced	8%			2%	16%	161
% unemployment benefit	2.5%			0.8%	5%	161
% houses for rent	46%					161
Degree of urbanism (1–5, 1= highest urban areas)		2.74	1.56	1	5	161

Neighbors make for about 20% of the network. Furthermore, social activities are highest and activities towards enhancing the neighborhood are relatively rare. Neighborhood satisfaction is high: on average 5.8 out of 7 points.

22.3.4 Analytical strategy

The interest in this paper is not the correlation of trust and control within individuals, but within neighborhoods. This means that individual level measures have to be aggregated to the neighborhood. A straightforward procedure of aggregation, such as a sum-score or a multiplication, however, would also aggregate the measurement error and would not take into account that there are differences between measuring individual level characteristics and collective ones. Individual level measurements obviously depend on characteristics of the individual. For example, older people might score systematically higher on trust (which they in fact do) or younger people might systematically rate a neighborhood as not close (which they also in fact do). In these cases, neighborhood composition determines answer patterns which ideally need to be filtered out in order to establish the 'true' level of control and trust. Furthermore, the items that measure control or trust are not independent of each other – just like residents in neighborhoods, these items are nested in respondents. To overcome these obstacles a procedure called 'ecometrics' has been established by Raudenbush and Sampson (1999, see also Mujahid et al. 2007; Raudenbush 2008). Ecometrics broadens the traditional psychometric assessment of individual traits, which usually distinguishes between levels – scale items nested within individuals – by adding a third level, the neighborhood. The goal is to establish a measurement of the aggregated level of the neighborhood. The procedure accounts for individual differences in responses to the given items and for the interdependency of these items. In a first step, a three-level model is estimated: neighborhoods, individuals, and individual level items measuring control and trust, respectively, and it is estimated how the different items contribute to the score on control and trust. Individual scores on neighborhood trust/control are estimated while controlling for characteristics that determine response patterns such as age, sex, migration background and education into account. Then, the residuals of that measurement, the parts that cannot be attributed to individual response patterns, constitute the measurement of trust/control at the aggregated level and it is used as the independent variable in the final analysis. Reliability (rho) of the ecometric based neighborhood measurements was .71 and .70 for control and trust, respectively. Pearson correlation between the simple aggregation of the values and the ecometric-based measure was .88 (control) and .87 (trust).

The final analyses comprise two levels and two kinds of models. For the number of neighbors in a network and the score for neighborhood satisfaction multilevel regression analyses (Snijders and Bosker 1999) were applied; the analysis for the actions with neighbors was a Poisson regression analysis (since values were

count data). Negative binomial models were also estimated, and they did not alter the conclusions (not reported here). All analyses started with the estimation of the effect of informal control expectations and then added the indicator of trust to the model. In all models, main effects and interaction effects of cohesion/trust and informal control are estimated.

22.4 Results

22.4.1 Control and trust

As described in the introduction to this paper, the scale for collective efficacy in the literature by Sampson and colleagues consists of cohesion/trust and informal control expectations. In their study, correlation between the scales for cohesion and control was .80, which justified to collapse them into one new scale, while arguing that the scales obviously were tapping into the same construct (Sampson, Raudenbush, and Earls 1997: 920). However, in the SSND, correlation between both scales is only .51, which is commonly considered to be only a moderate connection. Note that the high correlation between the scales is not always replicated though commonly assumed; see Reisig and Cancino (2004) who suggested that the high correlation is likely to exist in large cities (such as, in their case, Chicago).

22.4.2 Number of neighbors in network

Table 22.2 summarizes the multilevel regression estimating number of neighbors in the network.

Table 22.2: Multilevel regression on number of neighbors in network.

	M1			M2			M3		
	Coef.	**SE**	**p**	**Coef.**	**SE**	**p**	**Coef.**	**SE**	**p**
Informal Control	.105	.050	.036	.036	.059	.545	.029	.060	.063
Cohesion	–	–	–	.147	.058	.012	.145	.058	.013
Control*Cohesion	–	–	–	–	–	–	−.018	.041	.655
Individual control variables									
Sex	−.090	.043	.035	−.094	.046	.040	−.094	.046	.040
Age	.313	.050	.000	.322	.054	.000	.321	.054	.000
Education	.138	.049	.005	.160	.053	.003	.161	.053	.003
Background: native	.018	.059	.760	.030	.065	.642	.029	.065	.648

Table 22.2 (continued)

	M1			M2			M3		
	Coef.	**SE**	**p**	**Coef.**	**SE**	**p**	**Coef.**	**SE**	**p**
Total network size	.070	.010	.000	.079	.011	.000	.079	.011	.000
Neighborhood control variables									
% above 65 years of age	–.016	.058	.771	–.034	.061	.568	–.035	.061	.560
% non-western foreigners	–.040	.096	.673	–.027	.105	.795	–.023	.106	.827
% on unemployment benefit	.133	.064	.038	.179	.073	.016	.178	.074	.016
% divorced	.070	.070	.316	.076	.075	.310	.078	.075	.301
% houses for rent	–.122	.085	.154	–.118	.090	.191	–.118	.090	.190
Population density	.068	.062	.274	.044	.066	.504	.046	.066	.049
Intercept	1.76	.118	.000	1.71	.127	.000	1.72	.127	.000
Var (cons)	.023	.039		.024	.044		.025	.044	
Var (res)	1.320	.077		1.350	.083		1.348	.083	
LL	–1061.75			–1168.66			–1061.65		
Wald (chi)	130.96	.000		125.02	.000		124.96	.000	

Note: n = 973 respondents in 161 neighborhoods, variables are standardized. ICC = 3.5-4-5%.

Model 1 shows that informal social control is affecting the number of neighbors in the network: more control is associated with larger neighbor networks. Next to this, we found that from the individual level control variables age and education are associated with larger networks in the neighborhood. From the variables on the neighborhood level, a higher percentage of unemployed seems to stimulate many contacts in the neighborhood. Model 2 includes trust in the model and shows that the effect of informal control disappears if cohesion is included. More trust in the neighborhood residents is positively associated with the size of networks in the neighborhood. The interaction term in model 3 is not significant.

22.4.3 Activities with neighbors

Tables 22.3a-c summarize the analyses on different types of collective action: social activities, action towards improving the neighborhood and political action. In Table 22.3a we find again that the influence of informal control disappears if trust is added to the model. In addition, the interaction term between control and trust is significant: The association between trust and social activities in the neighborhood is weaker in the presence of control. In Table 22.3b, concerning the regression on collective action towards neighborhood improvement, both trust and control are significantly related with

Table 22.3a: Multilevel poisson regression on social activities in neighborhoods.

	M1			M2			M3		
	Coef.	**SE**	**p**	**Coef.**	**SE**	**p**	**Coef.**	**SE**	**p**
Informal Control	.217	.052	.000	.083	.056	.143	.094	.056	.097
Cohesion	–	–	–	.281	.053	.000	.265	.054	.000
Control*cohesion	–	–	–	–	–		−.096	.048	.048
Individual control variables									
Sex	.039	.040	.321	.063	.041	.128	.066	.041	.109
Age	−.038	.046	.408	−.066	.049	.174	−.088	.050	.080
Education	−.015	.045	.732	−.023	.047	.617	−.023	.047	.625
Background: native	−.006	.061	.917	−.035	.063	.575	−.064	.062	.299
Neighborhood control variables									
% above 65 years of age	.062	.049	.214	.066	.050	.190	.066	.050	.190
% non-western foreigners	−.196	.096	.042	−.175	.103	.090	.089	.106	.401
% on unempl.benefit	.022	.062	.715	.023	.060	.610	−.032	.073	.662
% divorced	−.133	.062	.715	−.009	.072	.206	−.094	.066	.152
% houses for rent	.001	.076	.988	−.081	.064	.886	.004	.079	.957
Pop. density	.042	.054	.436	−.011	.078	.885	−.012	.056	.822
Intercept	−.329	.044	.000	−.008	.056	.000	−.146	.114	.203
Var (cons)	.037	.007		.029	.047		.022	.013	
LL		−879.08			−786.17			−760.40	
Wald (chi)	87.77	.000	94.44		.000	66.11		.000	

Note: n = 973 respondents in 161 neighborhoods, variables are standardized. ICC = 3.5-4-5%.

Table 22.3b: Multilevel poisson regression on neighborhood related collective action.

	M1			M2			M3		
	Coef.	**SE**	**p**	**Coef.**	**SE**	**P**	**Coef.**	**SE**	**p**
Informal Control	.284	.069	.000	.183	.076	.016	.171	.077	.027
Cohesion	–	–	–	.213	.073	.003	.183	.074	.014
Control*Cohesion	–	–	–	–	–	–	−.045	.061	.462
Individual Control variables									
Sex	.077	.054	.157	.065	.056	.246	.089	.057	.117
Age	−.006	.062	.921	−.005	.064	.927	−.022	.067	.741
Education	.104	.064	.106	.117	.066	.076	.084	.068	.217
Background: native	.169	.093	.069	.119	.093	.203	.065	.093	.487
Neighborhood control variables									
% above 65 years of age	.018	.089	.838	.002	.089	.978	.006	.090	.944
% non-western foreigners	.074	.164	.650	.119	.166	.473	.245	.169	.148
% on unemployment benefit	−.061	.111	.583	−.090	.120	.450	−.117	.121	.333
% divorced	.023	.112	.837	.059	.114	.602	.069	.116	.055

Table 22.3b (continued)

	M1			M2			M3		
	Coef.	**SE**	**p**	**Coef.**	**SE**	**P**	**Coef.**	**SE**	**p**
% houses for rent	–.016	.136	.901	–.015	.136	.912	–.027	.138	.840
Population density	.062	.102	.546	.030	.103	.766	.027	.105	.795
Intercept	–1.01	.089	.000	–.919	.089	.000	–1.058	.176	.000
Var (cons)	.391	.104		.365	.103		.382	.107	
LL		–750.01			–693.39			–672.71	
Wald (chi)	33.70			.000	33.88	.007	.007	28.35	.000

Note: n = 973 respondents in 161 neighborhoods, variables are standardized. ICC = 3.5-4-5%.

Table 22.3c: Multilevel poisson regression on political collective action in neighborhoods.

	M1			M2			M3		
	Coef.	**SE**	**p**	**Coef.**	**SE**	**p**	**Coef**	**SE**	**p**
Informal Control	.255	.108	.018	.125	.116	.282	.159	.119	.180
Cohesion	–	–	–	.346	.108	.001	.362	.112	.000
Control*Cohesion	–	–	–	–	–	–	–.192	.108	.077
Individual control variables									
Sex	–.020	.023	.805	–.020	.084	.807	–.006	.084	.940
Age	.096	.095	.312	.071	.098	.469	.059	.102	.561
Education	.201	.096	.030	.217	.099	.029	.212	.101	.036
Background: native	.401	.185	.038	.361	.185	.051	.295	.182	.105
Neighborhood control variables									
% above 65 years of age	.136	.120	.255	.049	.123	.691	.046	.125	.709
% non-western foreigners	.107	.227	.635	.097	.226	.667	.222	.233	.341
% on unempl. benefit	.133	.145	.359	.207	.148	.164	.180	.154	.240
% divorced	–.174	.157	.264	–.106	.158	.499	–.110	.162	.499
% houses for rent	.048	.189	.799	.024	.185	.897	.028	.188	.879
Pop. density	.167	.137	.224	.134	.134	.980	.132	.139	.342
Intercept	–1.95	.137	.000	–1.87	.135	.000	–1.898	.262	.000
Var (cons)	.410	.169		.340	.154		.357	.160	
LL		–415.32			–396.55			–760.40	
Wald (chi)	26.61		.005	32.47		.001	25.50		.003

Note: n = 973 respondents in 161 neighborhoods, variables are standardized. ICC = 3.5-4-5%.

the outcome. The coefficient of control decreases, though, once trust is added to the model. There is no interaction effect. Table 22.3c shows, like the models in Table 22.3a, that the coefficient of informal control is not significant anymore in the presence of trust. In addition, the interaction between trust and control is negative: influence of cohesion is weakened by control.

22.4.4 Satisfaction with the neighborhood

Finally, Table 22.4 shows that both informal control and trust contribute to the satisfaction with the neighborhood (see model 2), although the coefficient of control decreases once trust is added to the model. In addition, the interaction between trust and control is found to be negative: the influence of trust is weakened in the presence of control.

Table 22.4: Multilevel regression on neighborhood satisfaction.

	M1			M2			M3		
	Coef.	**SE**	**p**	**Coef.**	**SE**	**p**	**Coef.**	**SE**	**p**
Informal Control	.314	.035	.000	.093	.033	.005	.039	.032	.227
Cohesion	–	–	–	.552	.033	.000	.497	.031	.000
Control*Cohesion	–	–	–	–	–	–	−.139	.022	.000
Individual control variables									
Sex	−.054	.031	.083	−.020	.026	.932	−.011	.024	.635
Age	.068	.039	.047	.002	.035	.720	−.221	.029	.456
Education	.069	.035	.052	.052	.030	.085	.049	.028	.086
Background: native	.017	.039	.658	.012	.035	.720	−.024	.035	.496
Neighborhood control variables									
% above 65 years of age	.042	.046	.362	−.001	.037	.974	.017	.033	.598
% non-western foreigners	−.376	.074	.000	−.318	.061	.000	−.239	.057	.000
% on unemployment benefit	.090	.049	.068	.117	.043	.007	.063	.040	.116
% divorced	−.093	.056	.100	−.025	.045	.581	−.013	.041	.738
% houses for rent	.043	.068	.524	.047	.045	.387	.051	.049	.293
Population density	−.091	.056	.074	−.154	.041	.000	−.140	.036	.000
Intercept	5.795	.037	.000	5.866	.030	.000	5.936	.069	.000
Var (cons)	.062	.024		.025	.016		.008	.012	
Var (res)	.716	.039					.395	.025	
LL	−1043.52			−766.49			−648.18		
Wald (chi)	253.16		.000	589.3		.000	124.9		.000

Note: n = 973 respondents in 161 neighborhoods, variables are standardized. ICC = 3.5-4-5%.

22.5 Intermezzo: Neighborhoods where control and trust coincide – or diverge

In a last step, the neighborhoods that score either high on one characteristic but low on the other or high/low on both are selected and via Google maps pictures are taken to explore and compare these neighborhoods. A low/high score is assigned if a neighborhood scores on control/trust at least one standard deviation lower/higher than the average. In 20% of the neighborhoods both the score of control as well as

of trust was relatively low. In 30% trust was low but control was high and in 15% the opposite held. In 35% of the neighborhoods, both trust and control were rated high. These 4 'types' of neighborhoods were examined and compared via google maps. Using this information, together with the information provided by Statistics Netherlands, I inductively searched for neighborhoods that represent their respective type. A typical neighborhood where both trust as well as informal control were found to be high is in small town, St Annaparochie in the Frisian part of the country. St Annaparochie lies close to the city of Leeuwarden and was founded in the 18th century in the course of the draining of the country and creating the polders. Many people who worked in the draining industry lived there and after the draining ended, they stayed in the new village. In addition, although it is very small – it has less than 5000 residents -, the village has good facilities, such as schools, sport centers and shops. An illustration for this neighborhood can be found here (https://tinyurl.com/y3uwyyhs).[7] The picture shows a neighborhood with detached single-family houses, seemingly built in the same time period, with gardens in front and probably also behind the houses. They look well-maintained and affluent, but at the same time not extraordinary rich. The streets seem quiet and there is much space and a lot of green.

A typical neighborhood where trust and control are both low is found in a street in Heemskerk in the neighborhood 'Oosterwijk', illustrated here (https://tinyurl.com/y2nv77oa). Heemskerk is located in the province North Holland, and belongs to a small municipality. Like St Annaparochie, it has been a rather rural area until in the sixties a rapid industrialization took place and the city became dominated by a big steal fabric (Hoogovens). Oosterwijk is precisely in between two municipalities, Heemskerk and Beverwijk. It is relatively homogeneous in terms of ethnic background – about 15% of the residents have a migration background. Incomes are relatively low, and the percentage of social housing relatively high. Note that the municipalities Heemskerk and Beverwijk differ in their social policies, which might constitute problems for the inhabitants of Oosterwijk. The picture shows a street with flat buildings, parking lots and a grocery shop. Buildings are probably from the 60s and there are many parking lots. Again, there is much space in the area, but houses are no single-family houses but flats for multiple households. There are no greenspaces and many of the small balconies are covered with a canvas sail, probably against the sun and heat. Last but not least there are signs of graffiti and there is an abandoned shopping car, probably belonging to the grocery but stalled around the corner.

7 Google maps permits links to street views, but no offline pictures, see the permissions (https://www.google.com/permissions/geoguidelines/) concerning geographic material. Hence, changes made by google may be reflected in the pictures. The advantage however is that the reader can use the links provided and examine the neighborhoods.

Where do control and trust not go together? A neighborhood were residents perceive high trust but low control has been found in Leiden (https://tinyurl.com/y5uqfqwa). Leiden is a relatively small town, close to the coast and surrounded by many tulip farms. Many houses are built in the 19th century or earlier, in medieval times. Leiden hosts the oldest and one of the bigger universities of the country, whereas it has only about 120,000 residents. Almost from the beginning, Leiden has been a city with a cleavage between the higher and lower educated and accordingly segregated neighborhoods. The picture shows a small alley, a bystreet typical for the inner city, with small low-rise houses, probably rented by students. The street is in the center of the city and there are backdoors from a shop or a restaurant, small signs of graffiti and not very well-maintained walls. Some windows are blinded, and many bikes are stalled the street is far too small for cars to pass.

Finally, a typical example of a neighborhood where trust is low, but control is high has been found in Den Haag, Mariahoeve, in the neighborhood 'Burgen en Horsten'. The neighborhood is close to 'Marlot', a neighborhood consisting of villa's and luxury apartments. Before its foundation in 1958, a train rode through the quite green area. The first houses that were built back then were small and meant for housing of blue-collar workers. In present days, about 50% of the residents is non-native and incomes are below the national average. The neighborhood consists in total of about 15,000 residents and is still a comparatively green neighborhood. Houses are often apartments for multiple households, often almost hidden behind trees. The illustration (https://tinyurl.com/yxzjsyqc) shows such a street with well-maintained garden in front of the houses. Windows are small for Dutch standards and the trees and green spaces seem to hide houses and people.

In terms of the traditional neighborhood characteristics, the four neighborhoods do not much differ in income and wealth. Leiden is the most expensive neighborhood; the housing price in the other three is below the national average (250.000 euro in 2019 according to Statistics Netherlands). Also, with the exception of the neighborhood in Leiden (8%), the percentage of elderly is between 21 and 25%. In Oosterwijk and Burgen en Horsten the percentage of migrants is relatively high (between 25 and 30 %), while it is low in the other two neighborhoods (between 1 and 10%).

This inductive exercise shows that describing the neighborhoods provides and intuitive understanding that the correlation between trust and control is area dependent. Importantly, we need more information and take more conditions into account when studying neighborhoods and their effects. History and embeddedness of the neighborhood itself seem at least as important as composition and wealth. Green spaces can indicate openness and invite for a chat, but they can also hide people and their housings from each other.

22.6 Conclusion and discussion

The analyses of the association of control and trust lead to a number of conclusions. First, control and trust are only moderately correlated in the Netherlands. This is an interesting finding that adds to the few studies that question the association between control and trust. Second, for the prediction of neighborhood functioning on an individual level, such as the number of neighbors in a network, collective actions with neighbors, and general satisfaction with the neighborhood, (perceived) trust is clearly much more important than control. Note that control has been measured as the expectation that residents will intervene and protect collective goods, if necessary. When included in the analysis, control at first seems to have an impact, but once trust is added to the model, effects of control become insignificant in all cases; only in the analysis of neighborhood satisfaction control remained having a (positive) influence. None of the analyses confirmed the idea that control and trust reinforce each other. However, interaction effects between control and trust turned out to be negative in the case of social activities and neighborhood satisfaction, indicating a substitution effect of trust. In the case of neighborhood collective action trust and control both have a positive main effect, so both contribute to actions towards neighborhood collective good production. This outcome is probably most in line with previous studies on collective efficacy.

It is important to note that neighborhood effects in the analyses were smaller than what is commonly found. The intraclass coefficients did not exceed 5%. This can be due to inappropriate neighborhood delineation but might also indicate that clustering in general is low. It is an important question for more comparative research to explain why the Dutch case seems to be so different from the American one.

When looking at the examples of neighborhoods where trust and control positively and negatively correlate a couple of interesting conclusions can be drawn – although this is material for further investigation, and, thus far, more illustrative or even anecdotical than established knowledge. Hopefully, future research will provide more arguments on this control-trust alignment by digging more systematically and theory driven than I did in this first account. First of all, it seems that trust and control are more likely to correlate positively in rural regions. In the case of St Annaparochie, it is a very small town, with a history of 'pioneering' residents. In the opposite example, in Heemskerk/Oosterwijk, where both control and trust have been found to be low, part of the explanation might be that industrialization has rapidly taken place, but also rural remnants remained. In a city such as Leiden and in a street, where obviously many young and highly educated people live, high trust but low control is reported. It seems that everyone is trusting, but monitoring is low and in situations where things have to be done and arranged, not-participation is not sanctioned. Finally, the case of Mariahoeve (Burgen en Horsten), an area built in the 60s for the working population is a case of a low income neighborhood where trust is not established, although one can

argue that all ingredients are present: the neighborhood is relatively green and houses are not too close to each other; they are even often detached.

The study has a number of limitations. First, the number of respondents per neighborhood varied between 2 and 24, which limits the reliability of the constructed measurements at the neighborhood level. Second, neighborhood satisfaction was in general quite high (5.8 on a scale from 1–7) and the standard deviation was with 1.1 quite low. Third, as already mentioned, we cannot draw causal conclusions from these analyses; the causality between for instance trust and the number of neighbors in the network can run in both directions (see also Lanfear, Matsueda, and Beach, 2020, for a critical assessment of causality in collective efficacy studies). Fourth, another limitation of this study is that I did not inquire into neighborhood variation in the models, random slopes for neighborhoods were not included.

Despite of these limitations, the study showed that control and trust can operate independently of each other. This finding is in line with earlier empirical (see Reisig and Cancino 2004, Volker et al 2015) as well as theoretical accounts (see Cook, Hardin and Levi 2005) and indicates that more theorizing is needed in order to understand under which conditions control and trust correlate or not. In addition, the series of exemplary photographs showed that neighborhoods – at least in the Netherlands – could not be grasped by common dimensions that neighborhood researchers use such as residential fluctuation, poverty, and the number of foreigners. 'Classical' compositional indicators as well as indicators of the built environment seem to tell a limited story. Green spaces can integrate people as well as separate them. A middle-class neighborhood can show low trust and cohesion, although all the ingredients for high trust are present. Likewise, a neighborhood where many students and academics live, can be cohesive and trustful, but nobody would step forward and take action necessary for collective good creation. Future research might gain by focusing more on these empirical anomalies, not yet predicted by established theoretical arguments.

Last but not least, the history of a neighborhood – rapid industrialization, pioneers who drained land from the water, or a newly built neighborhood for people, who are at the lower end of the income ladder, all matter for trust and control. Neighborhoods have a 'memory' – in the sense that people do not move out altogether. They were built with a purpose, which probably keeps influencing neighborhood affairs. As such, knowledge about history and architecture/land-use can help to develop a better understanding of neighborhoods in the Netherlands.

To conclude, in the Netherlands, trust and control in neighborhoods do not go systematically together when explaining neighborhood matters. Outcomes seem more driven by trust and its aggregate, social cohesion, than by informal control. However, the relationship between trust and control in neighborhoods varies a lot and future research has to establish more systematically the contextual conditions under which control and trust align or substitute each other when explaining neighborhood collective action and networks.

References

Bandura, Albert. 1997. *Self-efficacy: The exercise of control*. London, Macmillan Publishers.

Bandura, Albert. 2000. "Exercise of human agency through collective efficacy." *Current Directions in Psychological Science*, *9*, 3, 75–78.

Bellair, Paul, E. and Christopher, R. Browning. 2010 "Contemporary disorganization research: An assessment and further test of the systemic model of neighborhood crime." *Journal of Research in Crime and Delinquency*, *47*(4), 496–521.

Bijlsma-Frankema, Katinka and Ana Cristina Costa. 2005. "Understanding the trust-control nexus." *International Sociology*, 20, 3:259–282.

Blau, Peter, M. 1964. *Exchange and Power in Social Life*. New York, John Wiley.

Bursik, Robert, J. 1988. "Social disorganization and theories of crime and delinquency: Problems and prospects." *Criminology*, *26*(4), 519–552.

Bursik, Robert, J. 1999. "The informal control of crime through neighborhood networks." *Sociological Focus*, *32*,1, 85–97.

Buskens, Vincent. 1999. *Social Networks and Trust*. Amsterdam: Thela Thesis.

Buskens, Vincent, and Werner Raub. 2002. Embedded trust: Control and learning. *Advances in Group Processes*, 19, 167–202.

Browning, Christopher, R., Robert, D. Dietz, and Seth, L. Feinberg. 2004. "The paradox of social organization: Networks, collective efficacy, and violent crime in urban neighborhoods." *Social Forces*, 83,2, 503–534.

Coleman, James S. 1988. "Social capital in the creation of human capital." *American Journal of Sociology* 94: S95–S120.

Coleman, James, S. 1990. *Foundations of Social Theory*. Boston, Belknap Press.

Cook, Karen, Margaret Levi, and Russel Hardin. 2005. *Cooperation without trust?* New York, Russell Sage Foundation.

Cummings, Larry L., and Philip Bromiley. 1996. "The organizational trust inventory (OTI)." *Trust in organizations: Frontiers of theory and research* 302, 330: 39–52.

Dekker, Henri C. 2004. "Control of Inter-Organizational Relationships: Evidence on Appropriation Concerns and Coordination Requirements." *Accounting, Organizations, and Society* 29, 27–49.

Flache, Andreas. 1996. *The double edge of networks: An analysis of the effect of informal networks on cooperation in social dilemmas*. Amsterdam: Thesis Publishers.

Gau, Jacinta, M. 2014. "Unpacking collective efficacy: the relationships between social cohesion and informal social control." *Criminal Justice Studies* 27, 2, 210–225.

Gambetta, Diego.1988. "Can we trust trust?" In: Diego Gambetta (ed.), *Trust: Making and Breaking Cooperative Relationships*, pp 213–237. Cambridge, MA, Blackwell.

Goddard, Roger D., Wayne K. Hoy, and Anita Wollfolk Hoy. 2004. "Collective efficacy beliefs: Theoretical developments, empirical evidence, and future directions." *Educational researcher*, 33,3, 3–13.

Goddard, Roger, D. 2001. "Collective efficacy: A neglected construct in the study of schools and student achievement." *Journal of Educational Psychology*, 93,3, 467.

Guseva, Alya, and Akos Rona-Tas. 2001. "Uncertainty risk and trust: Russian and American credit card markets compared." *American Sociological Review*, 66: 623–646.

Handy, Charles. 1993. *"Understanding Organizations."* 4th ed. London, Penguin.

Inkpen, Andrew C., and Currall, Steven C. 1997. "International Joint Venture Trust: An Empirical Examination", in Paul W. Beamish and Peter Killing (eds) *Cooperative Strategies: Volume 1. North American Perspectives*, pp. 308–34. San Francisco, CA, New Lexington Press.

Janowitz, Morris. 1991. "*On social organization and social control*." Chicago, IL, University of Chicago Press.

Lanfear, Charles C., Ross, L. Matsueda, and Lindsey R. Beach. 2020. "Broken Windows, Informal Social Control, and Crime: Assessing Causality in Empirical Studies." *Annual Review of Criminology* 3,1 doi.org/10.1146/annurev-criminol-011419-041541 In advance first posted online on October, 7, 2019.

Malhotra, Deepak, and J. Keith Murnighan. 2002. "The effects of contracts on interpersonal trust." *Administrative Science Quarterly* 47 (3), 534–559.

Marsden, Peter, V. 1987. "Core discussion networks of Americans." *American Sociological Review*, 52,1, 122–131.

Morenoff, Jeffrey D., Robert J. Sampson, and Stephen, W. Raudenbush. 2001. "Neighborhood inequality, collective efficacy, and the spatial dynamics of urban violence." *Criminology*, 39, 3, 517–558.

Mujahid, Mahasin S., Ana Diez-Roux, Jeffrey, D. Morenoff, and Trivellore Raghunathan. 2007. "Assessing the Measurement Properties of Neighborhood Scales: From Psychometrics to Ecometrics." *American Journal of Epidemiology*, 165:858–67.

Mulder, Laetitia, Eric van Dijk, David de Cremer, Henk A.M. Wilke. 2003. "Undermining trust and cooperation: The paradox of sanctioning systems in social dilemmas." *Journal of Experimental Psychology*, 42, 147–162.

Raudenbush, Stephen. 2008. Many Small Groups. In: *Handbook of Multilevel Analysis*, edited by Jan DeLeeuw and Erik Meijer, pp. 207–36. New York: Springer.

Raudenbush, Stephen and Robert Sampson. 1999. "Ecometrics: Towards a Science of Assessing Ecological Settings: With Application to the Systematic Social Observation of Neighborhoods." *Sociological Methodology* 29,1,1–41.

Reisig, Michael, D. and Jeffrey, M. Cancino. 2004. "Incivilities in nonmetropolitan communities: The effects of structural constraints, social conditions, and crime." *Journal of Criminal Justice*, 32, 1, 15–29.

Ring, Peter S., and Andrew. H. Van de Ven. 1994 "Developmental processes of cooperative interorganizational relationships." *Academy of Management Review*, 19:90–118.

Shaw, Clifford and Henry McKay. 1942. *Juvenile Delinquency and Urban Areas*. Chicago, IL, University of Chicago Press.

Salanova, Marisa, Susanna Llorens, Eva Cifre, Isabel M. Martínez, and Wilmar B. Schaufeli. 2003. "Perceived collective efficacy, subjective well-being and task performance among electronic work groups: An experimental study." *Small Group Research*, *34*,1, 43–73.

Sampson, Robert, J. and W. Byron Groves. 1989. Community structure and crime: Testing social-disorganization theory. *American Journal of Sociology*, 94 (4),774–802.

Sampson, Robert, J. and Stephen W. Raudenbush. 1999. Systematic social observation of public spaces: A new look at social disorder in urban neihghborhoods. *American Journal of Sociology*, 105, 3, 603–651.

Sampson, Robert J. 2006. "Collective efficacy theory: Lessons learned and directions for future inquiry." In: Francis T. Cullen, John Paul Wright, Kristie R. Blevins, (eds) *Taking Stock: The Status of Criminological Theory* 15: 149–67.

Sampson, Robert J., Stephen W. Raudenbush, and Earls, F. 1997. "Neighborhoods and violent crime: A multilevel study of collective efficacy. " *Science*, *277*(5328), 918–924.

Sampson, R. J., Jeffrey Morenoff., & Felton Earls. 1999. "Beyond social capital: Spatial dynamics of collective efficacy for children." *American Sociological Review*, 64 (5), 633–660.

Sampson, Robert J. 2003. "The neighborhood context of well-being." *Perspectives in Biology and Medicine*, *46*(3), S53–S64.

Snijders, Tom and Roel Bosker. 1999. *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling*. London: Sage.
SSND. 2014. *The survey of the social network of the Dutch. Wave 3*. Beate Volker and Gerald Mollenhorst. Data and Codebook. Utrecht University.
Statistics Netherlands. 2013/2014. *Key Figures of Neighborhoods*. The Hague: CBS.
St.Jean, Peter, K. B. 2007. *Pockets of Crime: Broken Windows, Collective Efficacy, and the Criminal Point of View*. Chicago: University of Chicago Press.
Vlaar, Paul W.L., Frans A.J. Van den Bosch, and Henk, W. Volberda. 2007. "On the evolution or trust, distrust, and formal coordination and control in interorganizational relationships. Towards and integrative framework." *Group & Organization Management*, 32, (4), 407–429
Volker, Beate, Henk Flap, and Siegwart Lindenberg. 2007. "When are neighbourhoods communities? Community in Dutch neighbourhoods." *European Sociological Review*, *23*(1), 99–114.
Volker, Beate, Gerald Mollenhorst, Wouter Steenbeek, Veronique Schutjens, and Henk Flap. 2015. "Lost letters in Dutch neighborhoods: A field experiment on collective efficacy." *Social Forces*, *94*,3, 953–974.
Warner, Barbara, D. and Rountree, Pamela, W. 1997. "Local social ties in a community and crime model: Questioning the systemic nature of informal social control.": *Social Problems*, *44*(4), 520–536.
Warren, Mark E. 1999 (eds). *Democracy and Trust*. Cambridge: Cambridge University Press.
Williamson, Oliver. E. 1975. *Markets and Hierarchies*. New York: Free Press.
Wilson, William, J. 1996. *When Work Disappears: The World of the New Urban Poor*. New York: Alfred Knopf.

Tom A.B. Snijders and Frank Kalter

23 Religious Diversity and Social Cohesion in German Classrooms: A Micro-Macro Study Based on Empirical Simulations

Abstract: The micro-macro transition is a core problem of sociological theory building. Micro-intentions and micro-behavior do not straightforwardly translate into corresponding phenomena on the macro level, due to potentially existing rival mechanisms and the dynamics and complexity of social interactions. This chapter proposes an integrated statistical approach to studying the micro-macro transition by combining a random coefficient multilevel approach with the Stochastic Actor-Oriented Model. This is elaborated for the substantively interesting and topical question whether the growing ethnic and religious diversity in our societies, along with the well-known tendency for homophily, necessarily lead to a decline in social cohesion. The German part of the CILS4EU data is used to tackle this empirically. We investigate how religious homophily plays out differently depending on the context defined by the composition of the classroom, and explore the potential of simulation methods to explain this macro-level phenomenon from micro-level network dynamics. The empirical puzzle as stated is answered by a model representing homophily in a straightforward way, taking account of the variability between classrooms and the uncertainty about the parameter values; but a closer analysis reveals a further puzzle, which we leave for future research.

23.1 Introduction

The growing diversity of Western societies, especially in ethnic and religious terms, has become a topic of major interest in social research. The seminal paper of Putnam (2007) has been especially influential, arguing that diversity is challenging the cohesion of modern societies. This hypothesis has stimulated an enormous, rapidly growing number of empirical papers speaking for or against this general claim (van der Meer and Tolsma 2014). The vast majority of the underlying analyses rely on standard

Note: We are grateful to Sebastian Pink and David Kretschmer for their work on the data set, and to them and also Christian Steglich for discussions and comments. We also thank a reviewer for helpful comments on an earlier version.

Tom A.B. Snijders, University of Groningen; University of Oxford
Frank Kalter, University of Mannheim

https://doi.org/10.1515/9783110647495-023

survey data, i.e., data sets with respondents as individual cases. As a rule, an individual-level variable, such as generalized trust, support for the welfare state, civic engagement, etc., is chosen as an indicator for social cohesion, and some variant of the regression approach is applied to estimate whether diversity as a context-level variable, often measured by the Herfindahl index, has an effect on the individual outcome variable controlling for other independent variables.

This research paradigm has certainly led to many helpful insights, the most obvious being that results are mixed, thus showing that the general claim is conditional and that careful differentiations are needed. Nevertheless, the dominance of the survey-based regression approach is somewhat surprising: social cohesion is a macro-level phenomenon, and applying regressions in the sketched way implicitly assumes that the macro phenomenon can simply be derived as a statistical aggregate of the individual outcomes. However, it has long been emphasized that the micro-macro transition is a core challenge in sociological theory building, and is often far from trivial. The keyword in this context is 'emergence'. Micro-intentions and micro-behaviors do not straightforwardly translate to phenomena at the macro-level, due to the dynamics and complexity of social interactions. This has been widely stressed by sociologists (among many others, Raub 1984; Hedström and Swedberg 1998; Raub et al. 2011; Kalter and Kroneberg 2014).

Two classes of tools are especially suited to express social dynamics in the empirical micro-macro transition. On the one hand, network-analytical tools explicitly represent the dynamics of social interactions (e.g., Snijders 2001; Stokman and Vieth 2004). Macro-level characteristics of social networks can directly capture the idea of social cohesion (Moody 2001; Kalter 2016; Kalter and Kruse 2014). Network analysis allows to model diverse mechanisms of social interaction producing macro-level phenomena, and to test the validity of these mechanisms empirically. Next to this, agent-based modelling has become a major workhorse in dealing with phenomena of emergence. An agent-based model (ABM) is a model of individual actors interacting with each other and with environmental constraints over time (Epstein 2006). ABMs are programmed in computer language and analysed inductively: by iterating the assumed agent behaviour in the context of many other agents dynamically over time, ABMs allow to investigate the macro-level consequences of this interaction (Macy and Willer 2002; Manzo 2010).

Stochastic actor-oriented models (SAOM) for network dynamics (Snijders 2001; Steglich et al. 2010; Snijders and Steglich 2015) combine both of these tools. Basically, they are agent-based models that make assumptions about the behavior of actors, foremost their choices of building ties to other actors. In addition, they incorporate elements of generalized linear statistical models and confront the assumptions with empirical data. Correspondence with observed network-level descriptives allows to assess the goodness of fit of model assumptions, and to estimate parameters determining the behavior of actors from empirical data.

In this paper we use the SAOM as implemented in the SIENA software (Ripley et al. 2020) in an integrated empirical approach to the micro-macro transition. We study the impact of religious diversity on social integration in classrooms of adolescents, relying on the network data contained in the first two waves of the German part of the CILS4EU data (Kalter et al. 2013). We describe the general pattern of the relation between religious diversity and social cohesion and then employ simulation methods to figure out in how far they may be explained by various network formation mechanisms.

23.2 Theory and past research

'Social cohesion' is a key term in the social sciences, and as is the case for many key terms, it has been used in inconsistent and often vague ways. In empirical research it has been operationalized by trust, civic engagement, attitudes towards the welfare state, and many other concepts (van der Meer and Tolsma 2014; Schaeffer 2013). In a straightforward understanding, however, it refers to the social ties between the members of a community or society, and network analysis provides a lot of measures to give it a precise meaning (e.g., Scott and Carrington 2011). Whatever the precise measure, this view emphasizes that social cohesion is a macro-level result of individual-level processes of tie formation.

When asking why diversity in general, and religious diversity specifically, could have an impact on social cohesion in these terms, the most obvious reason certainly is homophily. One of the most robust findings in the social sciences is that people tend to have more ties to others who are similar to themselves (McPherson et al. 2001), and religious homophily is a well-known manifestation (e.g., Windzio and Wingens 2014; Cook et al. 2017). While the tendency to relate to similar others is partly a matter of the opportunity structure (Blau 1977), which has also been called 'baseline homophily' (McPherson et al. 2001, p. 419), empirical network analysis has shown that this tendency is strong also net of the mere availability of contacts. This is sometimes referred to as 'inbreeding homophily' and holds for quite a range of characteristics, e.g., age, sex, occupation, education, ethnicity (Moody 2001; Mouw and Entwisle 2006; Wimmer and Lewis 2010), and in the USA it has proved to hold also, and especially strongly, for religion (Cheadle and Schwadel 2012).

The mechanisms behind 'inbreeding homophily' are less clear and it is challenging to disentangle them in empirical analyses. The most obvious starting point to explain homophily is to trace them back to individual preferences.

Sharing characteristics reduces the cognitive and physical costs of communicating, anticipating and evaluating behavior, building mutual expectations, developing trust, etc. (Kossinets and Watts 2009). There is also support for the idea that similar others are found to be more attractive in appearance (e.g., Byrne 1971;

Huston and Levinger 1978). While religion is not a visible characteristic per se, wearing religious symbols – such as cross necklaces or headscarves – might be signals in the process leading to the creation of ties; most importantly, however, sharing religious world views seems, like psychological factors in general, particularly relevant for the costs and utilities of maintaining already existing ties (Felmlee et al. 1990). Religions may even contain explicit norms to prefer fellow believers.

Note that religion, in given contexts, is usually empirically correlated with other characteristics that may also foster a preference for similar people, most importantly ethnicity which also encompasses linguistic and other cultural elements, but also socio-economic class. Thus, preferences for social status, the same language, or other cultural aspects in the choice of social ties, can lead to religious homogeneity, and it will be empirically challenging to disentangle the true reasons (McPherson et al. 2001; Moody 2001). Note also that even within a given clear-defined opportunity structure, such as a classroom, the choice of friends may depend on reasons that are related to other, unrelated opportunity structures. Most obviously, in our case, students within the same classroom might also meet in their leisure time in religious places, like churches or mosques, and their friendship might predominantly arise from the time spent around these events. In contrast to 'availability effects' that arise from the opportunities in the context under investigation (i.e., the classroom), these kinds of additional opportunities arising in further organizational contexts have been called 'propinquity' effects (Wimmer and Lewis 2010).

Whatever the more detailed mechanism behind religious homophily, its existence would suggest that – net of additional mechanisms and in a straightforward aggregation of individual choices – social cohesion, vaguely defined for the moment, would decrease with religious diversity. Higher diversity, by definition, means a lower likelihood that two randomly chosen individuals share a characteristic; therefore, if this characteristic is associated with homophilous preferences, higher diversity is associated with a lower likelihood that they form a social tie.

However, the pattern of ties is not a mere aggregation of individual homophilous preferences, and there are a series of additional mechanisms determining tie formation. Some of these mechanisms might amplify the effects of homophily, others might counteract. The most prominent, and empirically most firmly established of these mechanisms are reciprocity (an early reference, in a German school context, is Delitsch 1900) and transitive closure (e.g., Davis 1970). Reciprocity means that the likelihood to choose someone as a friend is increased, if this person in turn has chosen oneself as a friend. To understand the deeper mechanism behind this tendency, Social Exchange Theory (Emerson 1976) is a fruitful starting point (also see Block 2015). Friendship is basically regarded as an investment, and mutuality helps to increase the expected rewards in relation to the costs. Transitivity denotes, loosely speaking, the phenomenon that a person is more likely to choose another as a friend, if there is a common third friend. Here standard explanations build on opportunity structure arguments, following classical ideas of Simmel (1950) or Granovetter (1973), on the one

hand, and on the more attitudinal mechanisms of classical Balance Theory (Heider 1948), on the other hand (again, see Block 2015). A first major attempt to figure out empirically how these more general network formation mechanisms influence the macro-micro-macro relation between diversity and social cohesion is formed by the analyses of Kalter and Kruse (2014). They use the first wave of the CILS4EU data, for all included countries. They study the consequences of ethnic diversity, as expressed by the Herfindahl index, on social cohesion measured in several ways: the density of the friendship network, the reachability (defined as average reciprocal geodesic distance) in the friendship network, and estimated coefficients for Exponential Random Graph Models. Basically, they find that, despite a clear and strong tendency for ethnic homophily, there is hardly any relation between ethnic diversity and social cohesion.

23.3 Data

We study the consequences of religious heterogeneity on social cohesion, using longitudinal network data from the German part of the Children of Immigrants Longitudinal Study in Four European Countries (CILS4EU) (Kalter et al. 2017). This comparative panel study was started in late 2010, interviewing adolescents that were about 14 years of age in wave 1. In the first step of the sampling process, schools were drawn from a nationwide list of schools that enroll students at this age; depending on the expected share of children of immigrants, schools were classified into four different strata and disproportionate stratified random sampling was applied, oversampling schools with higher proportions. In the second step, as a rule, two classrooms were randomly chosen within each school. In the third step, all students within this classroom were selected. The German sample of the wave-1 data comprises 144 schools, 271 classrooms and 5,013 students (Kruse and Jacob 2016). A sociometric module could successfully be administered in 267 of these German classrooms. It contains, among others, the nomination of the up to five best friends within the classroom. In wave 2, a year later, the same module could be applied again in 203 of these classrooms (Kruse et al. 2016). Our analysis is based on waves 1 and 2. We selected all classrooms where at least 10 students participated in wave 1, and likewise for wave 2. These were 140 classrooms.

23.4 Classroom cohesion

In our micro-macro study, the micro level is defined by the individual student, and the macro level by the classroom. We consider diversity with respect to religion, focusing on the main minority religion in Germany, Islam, and following a binary approach where the minority is defined by the Muslim students and the majority by all non-Muslim students. (Thus, we are imputing a common religion to all non-Muslims . . .)

We study how religious heterogeneity of the classroom affects social cohesion. The heterogeneity is reflected by the proportion of Muslim students in the classroom, denoted by p.

As mentioned above, network theory provides a rich variety of measures for social cohesion differentially emphasizing its various sub-aspects. Moody and White (2003) give a very sophisticated treatment. For the purposes of this paper we decided to work with a straightforward structural measure that is simple and easy to interpret. It is based on the *reciprocated degree*, defined as the number of reciprocated friendships of an individual. Because of our focus on majority and minority processes, we employ two measures of social cohesion in a classroom: the classroom average of the reciprocated degrees of its majority (non-Muslim) students, and the same average for its minority (Muslim) students. This ranges in the data from 0 to 5, and the majority as well as the minority have a mean of 2.3. This pays no attention to issues of integration and connectedness, but may be considered a basic measure that may be considered in future work together with other measures of structural cohesion such as studied in Moody and White (2003) and Kalter and Kruse (2014).

Figure 23.1 shows the scatter plot of the average number of reciprocal friendships per classroom in wave 2, for each religious category separately, in dependence on the proportion p of Muslims in the classroom, with a smooth approximation. The plot shows decreasing curves, not far from linear, for p up to 0.5. Perhaps they are increasing for greater proportions; but the number of classrooms there is low, and it is not clear from the plot whether this conclusion is warranted.

To further assess this relationship we conducted a bivariate regression with classrooms as cases, and as the two dependent variables the average reciprocated degrees for the two religious categories in wave 2. The independent variables were p and, as a control variable, the total number of students used in the analysis for this classroom – this was the maximum possible number of reciprocated friendships; its mean is 20.9. The analysis was done for the 111 classrooms with $p<0.5$ and having at least 10 respondents in both waves. The plot shows that in this range, the effect of p is approximately linear. The regression coefficients did not differ significantly between the two dependent variables. The bivariate regression was calculated using function gls in package nlme of the statistical system R.

Table 23.1 shows that there clearly is a decreasing effect of p in the range $0 \le p < 0.5$ (the interaction between p and religion was not significant). The expected average number of reciprocal ties in a classroom without any Muslims ($p=0$), and an average size of 20.9, is equal to 2.1. For $p=0.5$ this drops to 1.6 (and 1.5 for the Muslims in such a class). The drop from 2.1 to 1.6 is considerable.

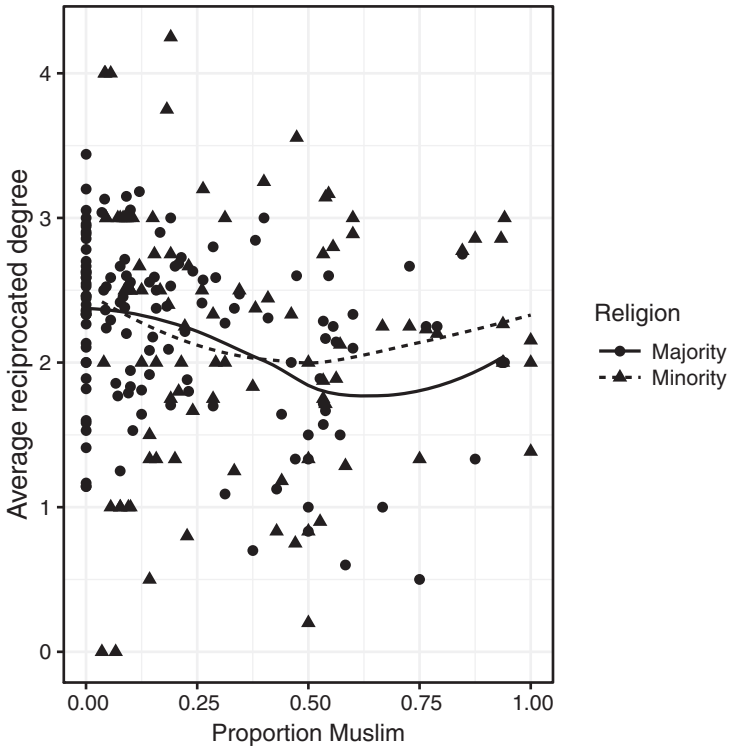


Figure 23.1: Average number of friendship nominations per classroom in wave 2, as a function of the proportion of Muslim students in the classroom, for majority and minority students.

Table 23.1: Bivariate regression of the average reciprocated degree, for minority and majority students, on the proportion minority students; *N*(classrooms) = 111.

Effect	**par.**	**(s.e.)**
Fixed part		
Intercept	1.796	(0.251)
Religion	–0.099	(0.114)
Proportion minority (*p*)	–0.961	(0.386)
Number of students with available data	0.015	(0.011)
Random part		
Variance majority	0.330	
Variance minority	0.603	
Correlation	0.418	

23.5 The empirical puzzle

Given that there is homophily to some extent with two groups, one being the majority and the other the minority, a straightforward expectation about the proportions in which the network is divided would be that, when the proportion of one group becomes larger, the average number of friends will become larger for this group and smaller for the other group, as a consequence of the availability of potential friends in the own group. This means that qualitatively and in a first approximation, we understand that in Figure 23.1 the curve for the majority is declining for minority proportions from 0 to 0.5, but not that for the minority, Muslims, it also is declining.

The empirical question therefore is twofold.

1. *Why is the average number of within-classroom reciprocated friendships for non-Muslim students declining as a function of the proportion of Muslims, for proportions less than 0.5? Can we understand the numerical value of this decline?*
2. *Why is the average number of within-classroom reciprocated friendships for Muslim students declining as a function of the proportion of Muslims, for proportions less than 0.5? Can we understand the numerical value of this decline?*

For the students in the majority group, our intuitive reasoning already seems to provide an answer to the first question; but we would like to back up this intuition with a formal empirical model, and study numerically the size of this decline. This is done in the next section. For this we use the Stochastic Actor-oriented Model, a model making the macro-micro-macro connection explicit. This will also serve as the starting point for studying the second question.

To express our two questions, we use two coefficients: the regression coefficient of the average reciprocated degree on p, controlling for the number of participating students in the classroom, where averaging is across the majority students; and this regression coefficient where averaging is across the minority students. We used only the 111 classrooms with at least 10 respondents in wave 1 and also in wave 2, and with less than 50% Muslim students, to avoid ambiguity about the definition of the minority. For this set of classrooms, the regression coefficient for the minority students is −1.23; for the majority it is −0.94. The question now is whether we can find satisfactory micro-level models to explain these numbers.

23.6 Multilevel network analysis

A model that represents the dynamics of network choices by students, given the context composed of the classroom, the attributes of all its students, and the current state of the network, is the Stochastic Actor-oriented Model ('SAOM'), proposed in Snijders (2001) and further explained in Snijders et al. (2010) and Snijders (2017).

It is implemented in the RSiena package (Ripley et al. 2020). The 'macro' here is the small environment consisting of the classroom, the 'micro' is the student who makes friendship choices. In the SAOM as applied to this case the set of actors is composed of the students in one classroom and the network is the structure of all friendship ties between them. It is assumed that the actors have control over their friendship choices, i.e., their outgoing friendship ties. The model takes the first observation of the network (wave 1) as given, and the dependent variable is the network at the second observation, i.e., wave 2. It assumes that the change from one network observation to the next is the result of a large number of sequential small changes, so-called micro-steps. In a micro-step one of the actors makes a choice in which the options are to create one new friendship tie, to withdraw one existing friendship tie, or to leave the network unchanged. The 'current network' changes gradually as a result of the micro-steps, from the first observed network to the second observed network. The choice in the micro-step is made with probabilities according to a generalized linear model in which the explanatory variables, called 'effects', are functions of the current network structure and the attributes of the actors. These probabilities can be derived from a myopic stochastic optimization principle, in which each actor optimizes a linear combination of the effects, called the 'evaluation function', to which random disturbances are added; the optimization considers only the direct result of this choice, without further strategic considerations. Detailed specifications are in Snijders et al. (2010). The choice of the effects, just like any model statistical specification, depends on theoretical considerations and empirical fit. The SAOM is applied here to the network dynamics from wave 1 to wave 2 of the CILS4EU data.

Figure 23.2 is Coleman's diagram (Coleman 1990; Raub et al. 2011) for the case of our study: on the basis of the classroom composition we wish to explain social cohesion as measured by the average reciprocated degree.

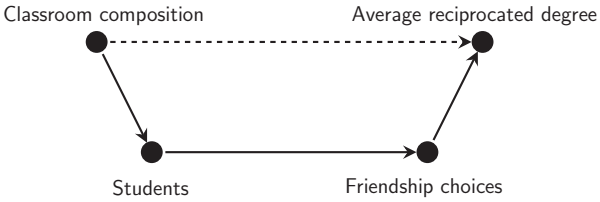


Figure 23.2: Coleman's scheme for this study.

The stochastic agent-based model at the core of the SAOM (Snijders et al. 2010; Snijders and Steglich 2015) is the basis of our macro-micro-macro approach. The bridge assumptions are represented by the specification of the SAOM, and how this depends on ethnicity and classroom composition; the assumptions about individual behavior are the probabilities of tie changes in the SAOM, which can be summarized

as myopic context-dependent goal-oriented behavior; and the transformation rule is the sequence of small changes (micro-steps) that takes the network from one observed wave to the next observed wave, each change also implying a change in the network context for all actors. The macro-level measure is a simple average of the actor-level reciprocated degrees, but it is only the end point of this quite complex transformation rule, and not a direct aggregate of individual choices.

We have quite a large set of German classrooms in the CILS4EU data, giving us ample variation in macro-level conditions. To handle the large number of classrooms, we need a multilevel version of the SAOM. Multilevel network analysis is discussed in Snijders (2016, p. 31–36). A multilevel SAOM is a combination of SAOMs, one for each classroom, all with the same model specification, but with possibly different parameters. To define the multilevel SAOM we have to specify how the parameters of the various classrooms are related, and how they are estimated. The simplest specification is the multi-group option (Ripley et al. 2020, Section 11.1), which assumes that all groups have the same parameter value. Even though this is quite a drastic assumption, it provides an illuminating first step in our micro-macro approach. A more reasonable approach is to assume that the parameters of the groups vary freely, and are estimated for each group separately. This approach resembles meta-analysis (Snijders and Baerveldt 2003). It was applied, e.g., in Knecht et al. (2010), where this approach posed some computational problems, because many of the school classes were too small to allow sound estimation by the methods implemented in RSiena; this was managed by a drastic reduction in the number of usable classrooms. The second step in this chapter follows a different approach: the integrated hierarchical multilevel approach developed in Koskinen and Snijders (2020), of which the implementation is described in Ripley et al. (2020, Section 11.3), as explained below.

23.6.1 First step: Multi-group approach

As a first step we try to answer our two questions while postulating that in each classroom the friendship network develops according to a Stochastic Actor-oriented Model with identical parameters across classrooms, but taking into account the different classroom compositions. As discussed above, we expect intuitively that the average reciprocated degree for students from the majority declines as a function of p, and increases for the minority. Therefore we do not expect that this initial model will provide the answer to both of our questions; but it is a check on the correspondence between our intuitive arguments and the Stochastic Actor-oriented Model.

The model specification was chosen parsimoniously and according to the current best practice. The effects included are explained in Snijders et al. (2010) and Ripley et al. (2020). The outdegree effect is like an intercept in other statistical models, and represents the balance between creating and dropping ties. Given that the

network is sparse, so potentially many more ties can be created than can be dropped, it usually has a negative parameter. Reciprocity and transitivity are basic features of network dynamics, represented by the reciprocity and transitive triplets effects. The 'transitive reciprocated triplets' is an interaction between reciprocity and transitivity, expected to be negative (Block 2015). Differential centrality of nodes is represented by the indegree-popularity, outdegree-activity, and outdegree-popularity effects; these reflect, respectively, variance of indegrees, variance of outdegrees, and the covariance between these. The 'reciprocal degree-activity' is the effect of the focal actor's reciprocated degree on tie creation and maintenance. This represents a 'saturation effect': an actor with a higher reciprocated degree is expected to have less value for additional ties, so that the parameter for this effect is expected to be negative. Sex homophily is usual in secondary school friendship networks, and represented by the 'different sex' effect. For religion, used here as a binary variable, the three basic effects are included, viz., the 'different religion' effect, the religion of the receiver (alter), and the religion of the sender (ego) of the friendship tie. The 'different religion' effect is expected to be negative, which represents the basic assumption in our micro-macro model, that being of a different religion has a negative effect on friendship creation and maintenance.

For the multi-group analysis, to achieve good convergence of the estimation algorithm, we included classrooms according to a somewhat more stringent criterion than above. To the requirements that both waves should have at least 10 respondents and the proportion Muslims should be less than 0.5, we added the criterion that there are at least 12 students for whom the reciprocal degree in wave 2 is non-missing, and that these are more than 60% of the total number in their classroom. This left 101 classrooms from the 111 selected in the previous section. For this data set, the regression coefficient of the average reciprocal degree on the proportion of Muslim students in the classroom is −1.16 for the minority students, and −0.74 for the majority.

Estimating this model under the assumption of equal parameters across groups led to the parameter estimates in Table 23.2. The estimated parameter values are in line with what is usually found for friendship networks in secondary schools. In particular, we see clear evidence for homophily with respect to sex and religion.

To see the implications of this model for the macro level, we simulated 1,000 data sets for the combined 101 schools according to the model of Table 23.2. For each simulated data set we computed the two regression coefficients of interest, i.e., the effect of the proportion of Muslim students on the average reciprocated degree for minority and majority students separately. This procedure is similar to the goodness-of-fit procedure usual for SAOMs (Lospinoso and Snijders 2019), but now applied to the multi-group situation. Figure 23.3 shows the distributions of these regression coefficients in violin plots.

For the minority students most of the distribution is in the positive range, for the majority students most is in the negative range. This is in correspondence to

Table 23.2: Parameter estimates and standard errors for the multi-group estimation of the Stochastic Actor-oriented Model, $N = 101$ schools.

Effect	estimate	(*s.e.*)
outdegree (density)	−0.809	(0.150)
reciprocity	2.902	(0.175)
transitive triplets	0.510	(0.028)
transitive reciprocated triplets	−0.178	(0.045)
indegree – popularity	−0.005	(0.018)
outdegree – popularity	−0.019	(0.053)
outdegree – activity	−0.114	(0.020)
reciprocated degree – activity	−0.231	(0.043)
different sex	−0.236	(0.034)
different religion	−0.177	(0.044)
religion Muslim alter	−0.031	(0.048)
religion Muslim ego	−0.084	(0.051)

Notes: Convergence t ratios all < 0.06.
Overall maximum convergence ratio 0.14.

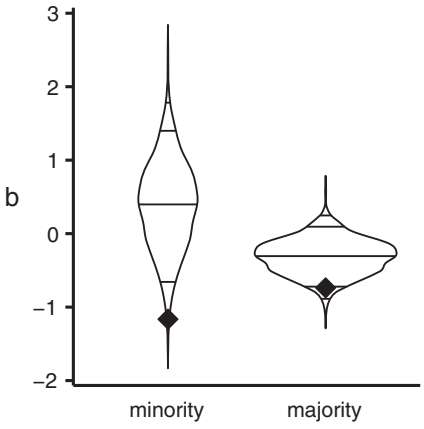


Figure 23.3: Distributions of regression coefficients of average reciprocated degrees on p for multigroup model. Horizontal lines denote quantiles at 0.01, 0.05, 0.50, 0.95, and 0.99; diamonds represent observed values.

our intuitive ideas that the availability of more potential friends in the own group will result in more reciprocated friendships; although the probabilities of these expected patterns are not very high. But we also see that the observed values are situated very low in the distributions; for the majority it is at percentile 0.10, for the minority at 0.01. The correspondence between the model predictions and the observed regression coefficient for the majority is a confirmation of our intuitive reasoning. However, this multi-group model does not correspond satisfactorily to our data with respect to the minority. Concluding, this model is sufficient to answer our first empirical question, but not the second.

23.6.2 Second step: Integrated multilevel approach

Could our questions then be answered, still by assuming that the model specification is identical between the classrooms, but the parameter values vary? This is more realistic than the previous approach. This is not expected to give systematically different average values for our regression coefficients, but there may be more random variability in a realistic way, which could imply that the data are not really as extreme as they seem to be in Figure 23.3.

The random coefficient multilevel version of the SAOM is developed in Koskinen and Snijders (2020). It extends the multilevel approach of Snijders and Baerveldt (2003) by assuming that the network in each classroom evolves according to a SAOM with the same specification, but different parameter vectors, these classroom-level parameter vectors having a multivariate normal distribution. This is similar to the Hierarchical Linear Model of multilevel analysis (Snijders and Bosker 2012), but now for longitudinal network data. A Bayesian estimation procedure for this model is implemented in the function sienaBayes of the R package RSienaTest (Ripley et al. 2020).

Convergence of the estimation posed some problems, and it was necessary to drop one classroom, because in attempts to estimate the model for the set of 101 classrooms it was an outlier, differing too strongly from the other classrooms. On inspection this classroom appeared to be composed of only majority students, with the highest average degree of all classrooms, 4.5. The regression coefficient of the average reciprocal degree on the proportion of Muslim students in the remaining set of 100 classrooms for the majority is −0.73. The conclusion for the multi-group approach for this set of 100 classrooms is the same as for the 101 classrooms analyzed above.

All SAOM parameters were assumed to vary randomly between classrooms. The macro-level parameters of the multilevel SAOM are the expected values and the between-classroom variances of the classroom-level parameters. Their estimates are presented in Table 23.3. Most of the estimates are similar to those in Table 23.2; the effects that are significant in both models, which are most, have the same sign. To the extent that there are differences, clearly the results from Table 23.3 are more credible, being based on more plausible model assumptions.

For this model, since it is Bayesian, we can use the so-called posterior predictive distribution (see, e.g., Jackman 2009) to check the implications and the model fit. In this case, this is the distribution where the set of 101 schools is fixed, as well as the composition of the classrooms and the friendship networks at wave 1, but the probability distributions are random: 'a sample from what the SAOM parameters possibly could have been'; and the networks at wave 2 also are random: 'a sample of what could have occurred in these schools given the sampled parameters'. The posterior predictive distribution reflect this double stochasticity. The posterior predictive distribution of our two regression coefficients, for the model of Table 23.3, are presented in Figure 23.4 again by violin plots.

Table 23.3: Posterior means and standard deviations for multilevel SAOM analysis for $N = 100$ classrooms.

Effect	par.	(*psd*)	betw. sd
outdegree (density)	−1.336	(0.105)	0.432
Reciprocity	2.338	(0.095)	0.278
transitive triplets	0.546	(0.025)	0.121
transitive reciprocated triplets	−0.199	(0.035)	0.158
indegree – popularity	0.044	(0.013)	0.087
outdegree – popularity	−0.118	(0.025)	0.130
outdegree – activity	−0.099	(0.017)	0.076
reciprocated degree – activity	−0.100	(0.026)	0.101
different sex	−0.320	(0.038)	0.205
different religion	−0.227	(0.064)	0.250
religion Muslim alter	0.038	(0.058)	0.212
religion Muslim ego	0.092	(0.072)	0.286

Notes: par = posterior mean; psd = posterior standard deviation;
betw. sd = posterior between-groups stand. deviation.

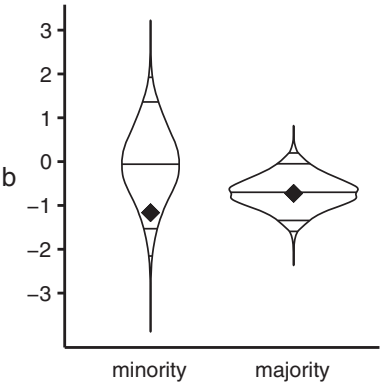


Figure 23.4: Distributions of regression coefficients of average reciprocated degrees on p for random coefficient model. Horizontal lines denote quantiles at 0.01, 0.05, 0.50, 0.95, and 0.99; diamonds represent observed values.

As expected, the difference with Figure 23.3 is mainly that the distribution in Figure 23.4 is more spread out (note the scale difference on the vertical axis). For the majority students the observed value is quite in the middle of the distribution (at percentile 0.47), and for the minority students it is still on the low side but not exceptional (at percentile 0.08). This means that for the majority as well as for the minority students, in the light of the variability between classrooms and the regular friendship network processes, there is nothing surprising about the observed negative regression coefficients. We may consider that both our questions can be answered by the SAOM, assuming parameter values to vary between the classrooms, and given the uncertainty that we have about their distribution.

23.6.3 And yet . . .

Our empirical puzzle focused on the regression coefficients for the average reciprocated degree on the proportion of minority students, and we answered it by considering posterior predictive checks for our macro-micro-macro statistical model. It should be noted, however, that such model checks are not an overall test of goodness of fit for the model. Posterior checks for statistics that are chosen for their descriptive interest may be rather forgiving, because such statistics may include a lot of variability that is not of diagnostic interest for the model. For example, the regression coefficients used depend also on the standard deviations of the average reciprocated degrees.

Therefore, for the purpose of model checking, we also consider the fit for a statistic that borrows less variability from elements of the model that are not of primary interest. Such a statistic is the partial correlation between the average reciprocated degree and the minority proportion, controlling for the number of cases and for the average reciprocated degree in the first wave. This partial correlation in our data set is −0.55 for the minority students, and −0.03 for the majority students. Figure 23.5 shows that this value is far from being reproduced by the random coefficient model of Table 23.3 for the minority students, although it is totally in line with respect to the majority students. If we had formulated our empirical puzzle in terms of this partial correlation coefficient, this random coefficient model would not have solved it.

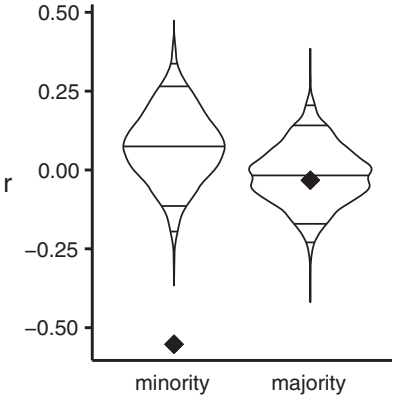


Figure 23.5: Distributions of partial correlations of average reciprocated degrees with p for random coefficient model. Horizontal lines denote quantiles at 0.01, 0.05, 0.50, 0.95, and 0.99; diamonds represent observed values.

23.7 Summary and conclusion

In this chapter we have tried to contribute to studying a key sociological question with an explicit notion of the micro-macro challenges involved. The macro-level consisted of the classroom (still small for macro), and there were a large number of macro-level cases. The micro-level consisted of the students. The research was about the dynamics

of friendship networks between the students measured at two waves, using the CILS4EU data. We have built on Knecht, Snijders, Baerveldt, Steglich and Raub (2010) where a similar multilevel network study was conducted, but without the explicit micro-macro focus; on Kalter and Kruse (2014) where a micro-macro study using the CILS4EU data was done, but without explicitly using network dynamics for the micro-level model; on Snijders and Steglich (2015), a micro-macro study using the SIENA framework, but not guided by a clear empirical puzzle; and on Koskinen and Snijders (2020), a multilevel network model.

Taking the relation between religious diversity and social cohesion as the substantive example, we used fundamental assumptions about network dynamics as the micro-level mechanisms, and tested to which extent these are sufficient to understand a particular macro-level characteristic. For the empirical test, we used rich multi-level longitudinal network data and recent elaborations of the Stochastic Actor-oriented Model. This enabled us to pursue an integrated approach to the macro-micro-macro question. An empirical puzzle, challenging common theoretical intuitions, served as the guideline in successively increasing the level of detail in the theoretical argumentation and the statistical implementation.

The puzzle consisted of two parts. We found that the regression coefficient of the classroom average reciprocated degree on the proportion of minority students is negative for the majority; and that it is also negative for the minority. Given a tendency to homophily, the first empirical result agreed with our intuitive expectations, and we expected no problems to confirm the empirical finding that this coefficient is negative. The second, however, was more puzzling. When the own group (minority) is larger and there is homophily, one would expect that the number of reciprocated friendships will be larger.

We approached these questions by applying a Stochastic Actor-oriented Model, specified in the usual way, assuming homophily. We focused not only on the signs, but wanted to find models giving a good correspondence with the values of the regression coefficients. When the assumption was that the parameters of the network model are identical across all classrooms, the regression coefficient found for the majority was well explained; but not so for the minority. However, after relaxing this auxiliary assumption by assuming that the classroom-level parameters are a sample from a multivariate normal distribution, also the coefficient for the minority was not unexpected at all; although, confirming our intuition, the estimated probability of a positive coefficient in our model was still larger than 0.5.

Therefore, one might say the puzzle is solved. However, the puzzle reappeared when we considered a less forgiving statistic: the partial correlation between average reciprocated degree and minority proportion, controlling for the average reciprocated degree at the earlier wave, is for the minority much lower than what could be reproduced by our Stochastic Actor-oriented Model with varying parameters.

Our puzzle was solved but the solution revealed a new puzzle, on which we hope to work in future. However, we did make substantive progress, and we hope

to continue with further theoretical and statistical enrichments. We do not regard this paper as the last tale on the substantive issues. Rather we understand it as a demonstration of the general fruitfulness of this approach, which comprises the collection of rich and demanding data and sophisticated elaborations of statistical modelling to detect theoretical desiderata, and which we hope will improve our understanding of micro-macro phenomena in the social world.

References

Blau, Peter M. 1977. *Inequality and Heterogeneity: A Primitive Theory of Social Structure*. New York: Free Press.

Block, Per. 2015. "Reciprocity, transitivity, and the mysterious three-cycle." *Social Networks* 40: 163–173.

Byrne, Donn. 1971. *The Attraction Paradigm*. New York: Academic Press.

Cheadle, Jacob E. and Philip Schwadel. 2012. "The 'friendship dynamics of religion,' or the 'religious dynamics of friendship'? A social network analysis of adolescents who attend small schools." *Social Science Research* 41:1198–1212.

Coleman, James S. 1990. *Foundations of Social Theory*. Cambridge/London: Belknap Press of Harvard University Press.

Cook, J. Benjamin, Philip Schwadel, and Jacob E. Cheadle. 2017. "The origins of religious homophily in a medium and large school." *Review of Religious Research* 59(1):65–80.

Davis, James A. 1970. "Clustering and hierarchy in interpersonal relations: Testing two graph theoretical models on 742 sociomatrices." *American Sociological Review* 35:843–852.

Delitsch, Johannes. 1900. "Über Schülerfreundschaften in einer Volksschulklasse." *Zeitschrift für Kinderforschung* 5:150–163.

Emerson, Richard M. 1976. "Social exchange theory." *Annual Review of Sociology* 2:335–362.

Epstein, Joshua M. 2006. *Generative Social Science: Studies in Agent-Based Computational Modeling*. Princeton, NJ: Princeton University Press.

Felmlee, Diane, Susan Sprecher, and Edward Bassin. 1990. "The dissolution of intimate relationships: A hazard model." *Social Psychology Quarterly* 53:13–30.

Granovetter, Mark S. 1973. "The strength of weak ties." *American Journal of Sociology* 78: 1360–1380.

Hedström, Peter and Richard Swedberg (eds.). 1998. *Social Mechanisms: An Analytical Approach to Social Theory*. Cambridge: Cambridge University Press.

Heider, Fritz. 1948. "Attitudes and cognitive organization." *Journal of Psychology* 21:107–112.

Huston, Ted L. and George Levinger. 1978. "Interpersonal attraction and relationships." *Annual Review of Psychology* 29:115–156.

Jackman, Simon. 2009. *Bayesian Analysis for the Social Sciences*. Chichester: Wiley.

Kalter, Frank. 2016. "Social network analysis in the study of ethnic inequalities." In *Emerging Trends in the Social and Behavioral Sciences*, eds. Robert A. Scott and Stephen M. Kosslyn. Oxford: Wiley, 1–15. doi:10.1002/9781118900772.etrds0397.

Kalter, Frank, Anthony F. Heath, Miles Hewstone, Jan O. Jonsson, Matthijs Kalmijn, Irena Kogan, and Frank van Tubergen. 2013. *Children of Immigrants Longitudinal Survey in Four European Countries (CILS4EU): Motivation, Aims, and Design*. Technical report, GESIS: GESIS Data Archive, Cologne.

Kalter, Frank, Anthony F. Heath, Miles Hewstone, Jan O. Jonsson, Matthijs Kalmijn, Irena Kogan, and Frank van Tubergen. 2017. *Children of Immigrants Longitudinal Survey in Four European* Countries *(CILS4EU) – Full Version*. Technical report, GESIS: GESIS Data Archive, Cologne. ZA5353 Data file Version 3.3.0, doi:10.4232/cils4eu.5353.3.3.0.

Kalter, Frank and Clemens Kroneberg. 2014. "Between mechanism talk and mechanism cult: New emphases in explanatory sociology and empirical research." *KZfSS Kölner Zeitschrift für Soziologie und Sozialpsychologie* 66:91–115.

Kalter, Frank and Hanno Kruse. 2014. "Ethnic diversity, homophily, and network cohesion in European classrooms." Pp. 187–207 in *Social Cohesion and Immigration in Europe and North America: Mechanisms*, Conditions, *and Causality*, eds. Ruud Koopmans, Bram Lancee, and Merlin Schaeffer. London: Routledge.

Knecht, Andrea, Tom A. B. Snijders, Chris Baerveldt, Christian E. G. Steglich, and Werner Raub. 2010. "Friendship and delinquency: Selection and influence processes in early adolescence." *Social Development* 19:494–514.

Koskinen, Johan H. and Tom A. B. Snijders. 2020. *Multilevel Longitudinal Analysis of Social Networks*. In preparation. Groningen/Melbourne.

Kossinets, Gregory and Duncan J. Watts. 2009. "Origins of homophily in an evolving social network." *American Journal of Sociology* 115:405–450.

Kruse, Hanno and Konstanze Jacob. 2016. *Children of Immigrants Longitudinal Survey in Four European* Countries. *Sociometric Fieldwork Report*. Technical report, Mannheim University, Mannheim. Wave 1-2010/2011, v1.2.0.

Kruse, Hanno, Markus Weißmann, and Konstanze Jacob. 2016. *Children of Immigrants Longitudinal Survey in Four European* Countries. *Sociometric Fieldwork Report*. Technical report, Mannheim University, Mannheim. Wave 2-2011/2012, v2.3.0.

Lospinoso, Joshua A. and Tom A. B. Snijders. 2019. "Goodness of fit for stochastic actor-oriented models." *Methodological Innovations* 12:2059799119884282.

Macy, Michael W. and Robb Willer. 2002. From factors to actors: Computational sociology and agent-based modeling." *Annual Review of Sociology* 28:143–166.

Manzo, Gianluca. 2010. "Analytical sociology and its critics." *European Journal of Sociology/ Archives européennes de sociologie* 51:129–170.

McPherson, Miller, Lynn Smith-Lovin, and James M. Cook. 2001. "Birds of a feather: Homophily in social networks." *Annual Review of Sociology* 27:415–444.

Moody, James. 2001. "Race, school integration, and friendship segregation in America." *American Journal of Sociology* 107:679–716.

Moody, James and Douglas R. White 2003."Structural cohesion and embeddedness: A hierarchical concept of social groups." *American Sociological Review* 68:103–127.

Mouw, Ted and Barbara Entwisle. 2006. "Residential segregation and interracial friendship in schools." *American Journal of Sociology* 112:394–441.

Putnam, Robert D. 2007. "E pluribus unum: Diversity and community in the twenty-first century." *Scandinavian Political Studies* 30:137–174.

Raub, Werner. 1984. *Rationale Akteure, institutionelle Regelungen und Interdependenzen. Untersuchungen zu einer erklärenden Soziologie auf strukturell-individualistischer Grundlage*. Frankfurt am Main: Lang.

Raub, Werner, Vincent Buskens, and Marcel A. L. M. van Assen. 2011. "Micro-macro links and microfoundations in sociology." *Journal of Mathematical Sociology* 35:1–25.

Ripley, Ruth M., Tom A. B. Snijders, Zsófia Bóda, András Vörös, and Paulina Preciado. 2020. *Manual for Siena Version 4.0*. Technical report, Oxford: University of Oxford, Department of Statistics; Nuffield College.

Schaeffer, Merlin. 2013. *Ethnic Diversity, Public Goods Provision and Social Cohesion: Lessons from an Inconclusive Literature*. Discussion Papers, Research Unit: Migration, Integration, Transnationalization SP VI 2013-103, WZB Berlin Social Science Center.

Scott, J. and P. J. Carrington (eds.). 2011. *The SAGE Handbook of Social Network Analysis*. London: Sage.

Simmel, Georg. 1917 [1950]. "Individual and society." Pp. 58–86 in *The Sociology of Georg Simmel*, ed. K. Wolff. New York: The Free Press.

Snijders, Tom A. B. 2001. "The statistical evaluation of social network dynamics." *Sociological Methodology* 31:361–395.

Snijders, Tom A. B. 2016. "The multiple flavours of multilevel issues for networks." Pp. 15–46 in *Multilevel Network Analysis for the Social Sciences; Theory, Methods and Applications*, eds. Emmanuel Lazega and Tom A. B. Snijders. Cham: Springer.

Snijders, Tom A. B. 2017. "Stochastic actor-oriented models for network dynamics." *Annual Review of Statistics and Its Application* 4:343–363.

Snijders, Tom A. B. and Chris Baerveldt. 2003. "A multilevel network study of the effects of delinquent behavior on friendship evolution." *Journal of Mathematical Sociology* 27:123–151.

Snijders, Tom A. B. and Roel J. Bosker. 2012. *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling* (2nd ed.). London: Sage.

Snijders, Tom A. B. and Christian Steglich 2015. "Representing micro-macro linkages by actor-based dynamic network models." *Sociological Methods & Research* 44:222–271.

Snijders, Tom A. B., Gerhard G. van de Bunt, and Christian E. G. Steglich 2010. "Introduction to actor-based models for network dynamics." *Social Networks* 32:44–60.

Steglich, Christian E. G., Tom A. B. Snijders, and Michael Pearson. 2010. "Dynamic networks and behavior: Separating selection from influence." *Sociological Methodology* 40:329–393.

Stokman, Frans N. and Manuela Vieth. 2004. "Was verbindet uns wann mit wem? Inhalt und Struktur in der Analyse sozialer Netzwerke." *Kölner Zeitschrift für Soziologie und Sozialpsychologie, Sonderheft* 44:274–302.

van der Meer, Tom and Jochem Tolsma. 2014. "Ethnic diversity and its effects on social cohesion." *Annual Review of Sociology* 40:459–478.

Wimmer, Andreas and Kevin Lewis. 2010. "Beyond and below racial homophily: ERG models of a friendship network documented on Facebook." *American Journal of Sociology* 116:583–642.

Windzio, Michael and Matthias Wingens. 2014. "Religion, friendship networks and home visits of immigrant and native children." *Acta Sociologica* 57:59–75.

Notes on the Editors and Contributors

Martin Abraham is Professor of Sociology and Empirical Social Research at the School of Business, Economics and Society at the Friedrich Alexander University Erlangen-Nürnberg, Germany. His research focuses on labor markets, organizations, economic exchange and households.

Ozan Aksoy is Lecturer (Assistant Professor) in Social Science at University College London, United Kingdom. His main research interests are cooperative behavior in modern societies and the political sociology of religion.

Marcel van Assen is Professor of Mathematical Sociology (special appointment by James Coleman Association) at the Department of Sociology/ICS, Utrecht University and Associate Professor at the Department of Methodology and Statistics and the Meta-Research Center, Tilburg University, the Netherlands. His research interests include mathematical sociology, social dilemmas, social networks, social science methodology, and meta-research.

Davide Barrera is Associate Professor at the Department of Culture, Politics and Society of the University of Turin and research affiliate at Collegio Carlo Alberto, Italy. His research interests are cooperation problems, behavioral game theory, experimental research, organizations, group processes, social networks, and actor-based social theory.

Michal Bojanowski is an Assistant Professor at the Chair of Quantitative Methods and Information Technology, Kozminski University, Poland. His research interests revolve around social networks and mathematical/computational social science as tools for understanding conflict and cooperation.

Vincent Buskens is Professor of Sociology at the Department of Sociology/ICS, Utrecht University, the Netherlands. His research interests include sociological theory, game theory, mathematical sociology, experimental sociology, social dilemmas, social networks, and institutions.

Rense Corten is Associate Professor at the Department of Sociology/ICS, Utrecht University, the Netherlands. His research interests include cooperation, trust, and (the dynamics of) social networks, with empirical applications including adolescent networks, social media, the sharing economy, online criminal networks, and laboratory experiments.

Andreas Diekmann is Professor em. of Sociology at the ETH Zurich, Switzerland (2003 – 2016) and a Senior Professor at the University of Leipzig, Germany (since 2018). His research interests focus on theories of social cooperation, experimental game theory, environmental and population sociology, and methods of empirical research.

Jacob Dijkstra is Associate Professor of Sociology at the Department of Sociology/ICS at the University of Groningen, the Netherlands. His research interests include experimental methods, game theory, social networks, and formal theory.

Christoph Engel is Director at the Max Planck Institute for Research on Collective Goods in Bonn, Germany, and Professor of Law at the Universities of Bonn and Rotterdam. His work centers on the behavioral analysis of legal intervention, mostly using experimental methods.

https://doi.org/10.1515/9783110647495-024

Hartmut Esser is Professor Emeritus of Sociology and Philosophy of Science at the University of Mannheim, Germany. His research interests include sociological theory, methodology of the social sciences, theories of action, migration, integration and ethnic conflicts, marital relations, and (currently) educational systems and educational inequality.

Andreas Flache is Professor of Sociology at the Department of Sociology/ICS at the University of Groningen, the Netherlands. His research focuses on computational modeling, social complexity, opinion dynamics, cooperation, experimental research, and social networks.

Henk Flap is Professor Emeritus of Sociology at the Department of Sociology/ICS at Utrecht University, the Netherlands. His research interests are social networks, the development of a social capital theory and testing it in various institutional contexts. Another of his research interests is the persecution of Jews in World War II.

Vincenz Frey is postdoctoral researcher at the Department of Sociology/ICS, University of Groningen, the Netherlands. His fields of interest include cooperation, trust, (the formation of) social networks, and processes of social influence.

Thomas Gautschi is Professor of Sociological Methodology at the University of Mannheim, Germany. His research and teaching interests include game theory, network analysis, model building, economic sociology, social science methodology and statistics, and experimental methods.

Francesca Giardini is Assistant Professor at the Department of Sociology/ICS, University of Groningen, the Netherlands. Her research interests include the mechanisms of social sustainability, especially the role of reputation and gossip on cooperation in groups, communities and organizations, inter-organizational networks, opinion dynamics and agent-based modeling.

Rainer Hegselmann is Professor of Philosophy at the Frankfurt School of Finance & Management, Germany. His research interests include modelling and simulation of social processes, opinion dynamics, social epistemology, philosophy of science, moral philosophy and the history of analytical philosophy.

Frank Kalter is Professor of Sociology at the University of Mannheim and Director of the German Center for Integration and Migration Research (DeZIM-Institute), Berlin, Germany. His major research interests include migration, integration, interethnic relations and the formal modeling of social processes.

Ferry Koster is Associate Professor at the Department of Public Administration and Sociology, Erasmus University Rotterdam, The Netherlands. His research interests include the sociology of labor market, organizations and institutions, and focuses on topics such as innovation, employability and social risks.

Siegwart Lindenberg is Professor of Cognitive Sociology at the Department of Sociology/ICS, University of Groningen, and the Department of Social Psychology, Tilburg University, the Netherlands. His interests lie in the development, testing and application of theories of social rationality that deal with the influence of the social environment on social need fulfillment, norms, cooperative behavior and self-regulation.

Tanja van der Lippe is Professor of Sociology the Department of Sociology/ICS, Utrecht University, the Netherlands. Her research interests include the work family interface, time use and time pressure, and the position of men and women on the labor market in a comparative way.

Kerstin Lorek is currently Representative for the School Board at the School of Business, Economics and Society, Friedrich-Alexander University Erlangen-Nürnberg, Germany.

Axel Ockenfels is Professor of Economics at the University of Cologne, Germany, and Speaker of the University's Excellence Center for Social and Economic Behavior. His research interests include market design and behavioral economics.

Karl-Dieter Opp is Professor Emeritus at the University of Leipzig, Germany, and Affiliate Professor at the University of Washington (Seattle), United States. His fields of interest are social theory, political participation, social norms and institutions, and the philosophy of the social sciences.

Bernhard Prosch was Professor of Sociology at the Friedrich-Alexander University Erlangen-Nürnberg, Germany. His main research areas were economic sociology, game theory, innovative methods of university teaching. He sadly passed away on September 23, 2015.

Arnout van de Rijt is Professor of Sociology at the European University Institute, Italy, and Utrecht University/ICS, the Netherlands. His research interests include computational social science, social networks, mathematical sociology, stratification, collective action, the sociology of science, the energy transition, immigrant integration, political polarization, and the spread of misinformation.

Anne Roeters is a senior researcher at the Netherlands Institute for Social Research (Sociaal en Cultureel Planbureau), The Hague, the Netherlands. Her research interests include time use, family sociology and the interface between work and family life.

Gerrit Rooks is Assistant Professor of Human Technology Interaction at Eindhoven University of Technology. The Netherlands. His main research interests are in humans embedded in social context, networks of relations and institutions, and how social embeddedness affects innovation and performance.

Stephanie Rosenkranz is Professor of Microeconomics at the Utrecht School of Economics, the Netherlands. Her research interests are in strategic framing and endogenous preferences, sustainable decision making, and social and economic networks.

Vera de Rover (Wiedemann) is currently Consumer and Sensory Insights Expert and Quality Manager at Essensor B.V. Essensor is a sensory market research company, specialized in supporting the food and non-food sectors in optimizing and selling products using market research that employs the human senses of smell, taste, vision, hearing and touch.

Esther de Ruijter is managing partner at Arbeid Opleidingen Consult, Giessenburg, the Netherlands. Her research interests include education and the labor market, household decision-making, rational choice theory and trust.

Chris Snijders is Professor of the Sociology of Technology and Innovation at the Human-Technology Interaction group of Eindhoven University of Technology, the Netherlands. His research

interests include human and computer-based decision making, online behavior and measurement, human-data interaction, and behavioral research methods.

Tom Snijders is Emeritus Professor of Statistics and Methodology at the Universities of Groningen, the Netherlands and Oxford, United Kingdom and Emeritus Fellow of Nuffield College (Oxford). He works at the interface of sociology and statistical modeling, with a focus on dynamics of social networks and multilevel modeling.

Frits Tazelaar was Professor of Sociology at the Department of Sociology/ICS, Utrecht University, the Netherlands. His research interests included relations within and between organizations. He sadly passed away on March 9, 2018.

Wout Ultee is Emeritus Professor of Sociology at the Department of Sociology/ICS, Radboud University Nijmegen, the Netherlands. His research interests include educational heterogamy, religious assortative marriage, coupled careers, thinking in analogies versus rational choices, and problem selection in sociology.

Beate Volker is Professor of Human Geography at Utrecht University, the Netherlands. Her research interests include social capital theory, the development of social networks and relationships through time, and the conditions for social cohesion and community in neighborhoods.

Thomas Voss is Professor of Sociology at the University of Leipzig, Germany. His research interests include rational choice and game theory, philosophy of social science, and economic sociology.

Fabian Winter is Head of Research Group Mechanisms of Normative Change at the Max Planck Institute for Research on Collective Goods, Bonn, Germany. His research interests include social norms, social dilemmas, and experimental sociology.

Rafael Wittek is Professor of Sociology at the Department of Sociology/ICS, University of Groningen, the Netherlands, and Scientific Director of the transdisciplinary research and training program Sustainable Cooperation (SCOOP). His research interests include sustainable cooperation and societal resilience, economic and organizational sociology, and social network analysis.