Allein die Masse macht's nicht - Antwort auf die Replik von Merkel et al. zu unserer Kritik am Demokratiebarometer

Jäckle, Sebastian; Wagschal, Uwe; Bauschke, Rafael

Postprint / Postprint
Zeitschriftenartikel / journal article

Empfohlene Zitierung / Suggested Citation:
Jäckle, S., Wagschal, U., & Bauschke, R. (2013). Allein die Masse macht's nicht - Antwort auf die Replik von Merkel et al. zu unserer Kritik am Demokratiebarometer. *Zeitschrift für Vergleichende Politikwissenschaft : German Journal of Comparative Politics*, *7*(2), 143-153. https://doi.org/10.1007/s12286-013-0148-7

This is an accepted manuscript of an article published by Springer in Zeitschrift für Vergleichende Politikwissenschaft, Vol. 7, No. 2, 2013, pp. 143-153 available online at: https://doi.org/10.1007/s12286-013-0148-7

Allein die Masse macht's nicht – Antwort auf die Replik von Merkel et al. zu unserer Kritik am Demokratiebarometer

Sebastian Jäckle, University of Freiburg
Uwe Wagschal, University of Freiburg
Rafael Bauschke, University of Freiburg

Wissenschaftlicher Fortschritt ist auf diskursive Prozesse und kritisch-konstruktive Auseinandersetzungen angewiesen. Mit unserem Beitrag zum Demokratiebarometer (Jäckle et al. 2012) wollten wir einen solchen produktiven Diskussionsprozess anstoßen, indem wir mögliche Defizite und Verbesserungspotentiale dieses Demokratiemaßes herausarbeiteten. Wir haben es daher auch mit großer Freude zur Kenntnis genommen, dass Wolfgang Merkel und seine Kollegen auf unseren Artikel reagiert und ihre Sicht auf unsere Kritikpunkte in einer Replik dargelegt haben (Merkel et al. 2013). Allerdings scheint unsere Kritik doch primär auf- statt angeregt zu haben (Merkel et al. 2013, S. 75). Bekanntlich lässt sich über stilistisches Empfinden vortrefflich streiten, gleichwohl halten wir eine Rückbesinnung auf die Auseinandersetzung um den Messansatz selbst an dieser Stelle für zielführender. In dieser Hinsicht hat die Replik bisweilen erhellt, in weiten Teilen vermochten uns die vorgebrachten Punkte jedoch nicht dahingehend zu überzeugen, dass wir unsere Position bezüglich des Demokratiebarometers (DB) grundlegend zu ändern hätten. An mancher Stelle scheint es, als seien unsere Argumente nicht verstanden bzw. anders als intendiert interpretiert worden. Wir danken daher den Herausgebern der ZfVP, dass sie uns die Möglichkeit einräumen, in aller Kürze auf den Beitrag von Merkel et al. zu antworten, und uns damit ermöglichen, jenseits jeglicher verbaler Hochrüstung solche Missverständnisse zu klären. Die Gliederung unseres Beitrags orientiert sich erneut an den drei von Munck und Verkuilen (2002) vorgeschlagenen Ebenen der Konzeptionalisierung (1), der Operationalisierung (2) und der Messung (3).

1. Konzeptionalisierung

Merkel et al. fokussieren hier auf zwei unserer Kritikpunkte: erstens die Konstruktionslogik des DB und zweitens die Einordnung des Demokratiekonzepts als eher maximalistisch.

1.1 Die Beziehung der Grundprinzipien Freiheit, Gleichheit und Kontrolle

Wir zogen in der Diskussion der Konstruktionslogik einen Artikel von Victoria Kaina aus dem Jahr 2008 heran (Kaina 2008). Auch wenn sich besagter Beitrag nicht auf die jüngsten Veröffentlichungen zum DB bezieht, lassen sich die in ihm aufgeführten Kritikpunkte unserer Meinung nach weiterhin unverändert anbringen, da es in Bezug auf die grundlegende Konstruktionslogik des Demokratiemodells keine relevanten Veränderungen beim DB gegeben hat und – so die Replik von Merkel et al. – wohl auch nicht geben wird. Zwar stimmen wir nicht

mit sämtlichen Argumenten Kainas überein[1], aber ihren Hauptkritikpunkt bezüglich der Beziehung der drei Grundprinzipien Freiheit, Gleichheit und Kontrolle, erachten wir weiterhin für bedenkenswert. Das DB geht davon aus, dass diese in einem interdependenten Verhältnis zueinander stehen. Doch wie wird diese Interdependenz verstanden? Und am wichtigsten: Welche Auswirkungen hat dieses Verständnis auf die weitere Messung? Merkel et al. stellen in ihrer Replik klar, dass die wechselseitige Abhängigkeit von Fall zu Fall unterschiedlich verlaufen kann (Merkel et al. 2013, S. 76): In einem Fall mögen beispielsweise hohe Gleichheitswerte die Freiheit begrenzen, in einem anderen, genau umgekehrt, verstärken sich die beiden Werte gegenseitig.[2] Da wir es mit drei Grundprinzipien zu tun haben, die allesamt in einem solchen, potentiell in beide Richtungen funktionierenden interdependenten Verhältnis zueinander stehen, lässt sich das Demokratieverständnis des DB am besten als ein äußerst komplexes, dynamisches Gleichgewicht beschreiben, welches durch das jeweils fallspezifisch optimale Austarieren der drei Grundprinzipien entsteht. Die Autoren des DB lassen dabei ganz bewusst offen, welche Austarierungsoption und damit Verteilung zwischen den drei Grundprinzipien „optimal“ ist. Sie gehen vielmehr davon aus, dass es von Fall zu Fall verschiedene Optimierungsmöglichkeiten geben kann. Auf den Punkt gebracht: zwei Mal hoch und ein Mal niedrig könnte bei den Grundprinzipien genauso „optimal“ sein wie drei Mal mittelmäßig.

Stellt man diese Logik nun allerdings der tatsächlich verwendeten Aggregationsregel gegenüber, bleibt von der Konzeption von Demokratie als System interdependenter demokratischer Grundprinzipien nämlich nicht viel mehr übrig als die Annahme, dass eine Demokratie eine besonders hohe Qualität aufweise, sofern alle drei Prinzipien möglichst hoch ausgeprägt sind. Hier ist die Aggregationsregel also deutlich weniger komplex als es die grundlegende Konzeption erwarten ließe. Zusätzlich nimmt das DB entsprechend seiner Aggregationsregel eine (deutlich begrenzte) gegenseitige Substituierbarkeit unter den Grundprinzipien an. Durch die Aggregation mittels Arkustangens wird implizit dann eben doch eine bestimmte Verteilung der Grundprinzipien (und davor schon der neun Funktionen, die ebenfalls auf diese Weise zu den Grundprinzipien aggregiert werden) favorisiert: Je ausgewogener die zu aggregierenden Werte vorliegen, desto höher ist der aggregierte Wert.

[1] So halten wir den Punkt, dass es keinen optimalen Erfüllungsgrad der Demokratieprinzipien geben kann, da „ein Optimum *entweder* erreicht ist *oder* nicht“ (Kaina 2008, S. 522), für eine in erster Linie semantische Argumentation, die inhaltlich nicht sonderlich weiterführt.

[2] Gleichwohl dürfte der Fall der negativen gegenseitigen Beeinflussung, wie wir ihn in unserer ursprünglichen Kritik angenommen hatten, schon eher die Regel sein, insbesondere wenn man die Bereiche mit vergleichsweise höheren Merkmalsausprägungen betrachtet. Laut Bühlmann et al. ist nämlich „eine gleichzeitige Maximierung aller neun Demokratiefunktionen aufgrund eines gewissen Spannungsverhältnisses zwischen den Prinzipien ‚Freiheit‘ und ‚Gleichheit‘ kaum möglich“ (Bühlmann et al. 2012, S. 144).

Diese Aggregationslogik steht in unseren Augen entgegen der Grundannahme, dass komplett unterschiedliche Austarierungsoptionen zwischen den drei Grundprinzipien letztlich dieselbe Demokratiequalität bedeuten können.

1.2 Ein *zu* maximalistisches Demokratieverständnis

In unserem Beitrag haben wir das grundlegende Demokratiekonzept des DB als „eher maximalistisch[]“ (Jäckle et al. 2012, S. 107) bezeichnet, was von Merkel et al. kritisiert wird. Sie sehen ihr Konzept „in einer mittleren Position“ (Merkel et al. 2013, S. 77). Ungeachtet dieser Wortklauberei haben wir in unserer Kritik vor allem auf die Gefahren hingewiesen, die eine zu maximalistische Konzeption mit sich bringt: Da Merkel et al. zudem eine Begründung für unsere Einschätzung vermissen – obgleich sich diese in den folgenden Abschnitten unseres Artikels gefunden hätte – wollen wir hier die zwei wichtigsten Punkte nochmals kurz wiederholen und zusammenfassen.

Erstens bleiben wir bei unserer Einschätzung, dass das Konzept des DB eher zu viele als zu wenige Komponenten beinhaltet. Auf der obersten Ebene des Konzeptbaumes ist dies zwar noch nicht der Fall, aber bereits auf der direkt darunter liegenden Ebene der neun Funktionen lässt sich ein gewisses „conceptual stretching“, beispielsweise bei der Rechtsstaatlichkeit, ausmachen. Hier möchten wir auf einen Text von Gerardo Munck verweisen, den Merkel et al. in ihrer Replik als eine „positive[] Rückmeldung“ (Merkel et al. 2013, S. 75) auf das DB ausweisen. Zumindest bezogen auf die Probleme der konzeptionellen Überdehnung und der Abgrenzung der einzelnen konstituierenden Komponenten voneinander liest sich Muncks Text jedoch durchaus kritisch:

> “suggesting that the rule of law ([…], Bühlmann, Merkel, Müller and Wessels 2011: 6) […] is a conceptual component of the quality of democracy does little to advance the process of conceptualization. Since the rule of law, freedom and equality are such broad principles, their inclusion as part of the quality of democracy is fraught with problems. It is not only hard to tell if they should or should not be seen as part of the quality of democracy. Even if they do deserve to be considered a component of the quality of democracy, they are notoriously hard to distinguish from other components” (Munck 2012, S. 12).

Zweitens: Als besonderes Problem begreifen wir in Bezug auf die Konzeptionalisierung die schrittweise Ausdehnung des grundlegenden Konzepts über die einzelnen Stufen von den drei Grundprinzipien, über die neun Funktionen, 18 Komponenten, 51 Subkomponenten hin zu den 100 Indikatoren. Jeder dieser Schritte und auch die schiere Masse an Einzelindikatoren erhöhen unserer Meinung nach erstens die Gefahren der *redundancy* und *conflation* und zweitens die Gefahr, dass auf diese Weise zumindest teilweise *outputs* und/oder *outcomes* guter

Regierungsführung – wir hatten hier auf den Indikator niedrige Mordrate verwiesen[3] – als Indikator für die Demokratiequalität herangezogen werden – etwas, das die Macher des Demokratiebarometers eigentlich dezidiert zu vermeiden suchen (Bühlmann et al. 2012, S. 18).[4] Auch zu diesem Thema findet sich ein sehr aufschlussreicher Abschnitt in dem von Merkel et al. zitierten Papier Gerardo Muncks, welcher zudem in seiner zentralen Überblickstabelle, ähnlich wie wir, beim DB durchaus eine Reihe „intermediärer" wie „finaler" *outcomes* integriert sieht (Munck 2012, S. 10):

"The outcomes to which a substantive conception of democracy draws attention are all fundamental. But they deserve to be treated separately. Most fundamentally, including a long list of civil and social rights into a definition of democracy contradicts the essence of democracy as a set of procedures for determining the content of public policy. Indeed, just as an election is not deemed to be democratic because the 'right candidate' wins, a political system should not be deemed democratic only if it adopts the 'right policies.' Moreover, civil and social rights might be advanced without democracy and hence it is recommendable to break down the compound concepts of liberal democracy and social democracy and assess empirically how democracy compares to other political systems in the promotion of civil rights and social rights more broadly conceived. Indeed, it is crucial to avoid making democracy synonymous with all that might be considered good about politics." (Munck 2012, S. 34-35)

2. Operationalisierung

Im Hinblick auf die Operationalisierung nehmen Merkel et al. Stellung zu der von uns problematisierten mangelnden Detailliertheit in Bezug auf die Entwicklung des Konzeptbaums sowie der von uns kritisierten fehlenden Stringenz der Indikatorenauswahl im Hinblick auf das der Messung zugrunde liegende Konstrukt.

---

[3] An dieser Stelle hätten wir auch eine Antwort auf unsere viel allgemeinere Frage erhofft, inwiefern es ein Problem für die Messung demokratischer Qualität darstellt, wenn ein Indikator sowohl in ‚gut funktionierenden' autokratisch regierten Systemen als auch in ‚gut funktionierenden Demokratien' ähnliche Ausprägungen annimmt. Kann ein Indikator als Maß für die Qualität entwickelter Demokratien also problemlos verwendet werden, wenn er gleichzeitig keine Unterscheidung zu autokratischen Systemen ermöglicht? In ihrer Replik scheinen uns Merkel et al. hier nicht richtig verstanden zu haben. Keineswegs geht es uns darum, normativ die Qualität autokratischer Staaten über die in diesen vorherrschende Mordrate zu bestimmen, vielmehr sehen wir die Gefahr, dass hier ein Indikator herangezogen wird, bei dem fraglich ist, ob er als Teilmaß für demokratische Qualität wirklich geeignet ist. Zudem schreiben Merkel et al., dass unsere diesbezüglichen Überlegungen sowieso nicht zutreffen, da sie das DB „explizit und unübersehbar für Demokratien entwickelt haben" (Merkel et al. 2013, S. 79). Allerdings stellt schon Munck fest, dass die Autoren des DB „focus solely on ‚established democracies' in their empirical work but nonetheless suggest, if somewhat unclearly, that they could ‚measure degrees of democratic qualities even in not fully-fledged democracies' (Bühlmann, Merkel, Müller, Giebler and Wessels 2011: 9-10)" (Munck 2012, S. 15). Insofern fragen wir uns, welcher Schwellenwert es ist, der die von Merkel et al. zu bemessenden Demokratien von solchen teildemokratischen Staaten trennt, in denen die Mordrate wiederum eher als ein Ausweis gut funktionierender autokratischer Tendenzen interpretiert werden könnte?

[4] Allein der komparative Vergleich mit allen prominenten Demokratieindizes (Freedom House Index, Politiy IV, Bertelsmann Transformations Index BTI, Vanhanen Index, Democracy and Dictatorship (DD), Index of Democracy) sollte die Autoren etwas selbstkritischer werden lassen: Das Demokratiebarometer verwendet mit deutlichem Abstand die meisten Indikatoren.

2.1 Dokumentation und Herleitung der Indikatoren

Merkel et al. schreiben bezogen auf unsere Kritik an der Dokumentation und Herleitung der Indikatoren, dass „die Ableitung nun für 18 Komponenten, 51 Subkomponenten und 100 Indikatoren zu fordern, [...] für wissenschaftliche Publikationen witzlos“ (Merkel et al. 2013, S. 78) sei. Diese Aussage ist in unseren Augen wenig hilfreich. Wir bezogen unsere Kritik und Verbesserungsvorschläge auf das gesamte Projekt des DB und nicht nur auf einen Zeitschriftenartikel. Gerade auf der, im Übrigen von uns explizit positiv für ihren Aufbau und die Datendokumentation hervorgehobenen Projekthomepage wäre der Platz vorhanden, eine umfassende, über eine bloße Auflistung hinausgehende, schrittweise Herleitung aller verwendeten Indikatoren im Sinne des zugrunde liegenden Konzeptbaums sauber zu dokumentieren. Der Verweis darauf, dass andere Messungen eine solche Dokumentation auch nicht leisten, sollte an dieser Stelle wohl eher als Argument für und nicht gegen eine solche Arbeit dienen. Aus diesem Grund begrüßen wir auch sehr den Vorschlag, das Codebuch um die Ableitung sämtlicher Indikatoren zu ergänzen, denn dies ermöglicht es interessierten Wissenschaftlern für sich selbst zu entscheiden, ob die Aufnahme bestimmter Indikatoren das grundlegende Konzept in ihren Augen gut operationalisiert oder doch über Gebühr ausweitet.

2.2 Stringenz der Indikatorenauswahl

Merkel et al. monieren, dass wir unser Argument auf Basis von „sage und schreibe dreieinhalb Indikatoren“ (Merkel et al. 2013, S. 78) entwickeln. Hier sei zunächst darauf verwiesen, dass auch wir im Rahmen einer Zeitschriftenpublikation nur begrenzten Platz haben. Gleichwohl halten wir substantiell an den gewählten Indikatoren fest, da diese uns zur Veranschaulichung spezifischer grundlegender Probleme dienen, die wir in Bezug auf die Stringenz der Indikatorenauswahl beim DB sehen. Weder behaupten wir, dass diese Probleme für alle Indikatoren des Demokratiebarometers gleichermaßen gelten, noch hatten wir beabsichtigt einen Anteil *nicht-stringenter* Indikatoren zu bestimmen, um hierdurch beispielsweise die prozentuale Güte des DB zum Ausdruck zu bringen.

Der erste dieser umstrittenen Indikatoren ist die Mordrate, die uns zu zwei generellen Fragen führt: Erstens, ist eine Variable, die sowohl in gut funktionierenden demokratischen Ländern als auch in autokratischen Ländern mit gut funktionierenden Repressionsapparaten eine sehr ähnliche Ausprägung annehmen kann und damit diese nur mäßig diskriminiert, geeignet,

demokratische Qualität zu bemessen.[5] Und zweitens, das weiter oben bereits angesprochene Problem des Überdehnens des eigentlichen Konzeptes durch die Aufnahme von *policy outcomes,* zu denen wir die Mordrate zählen.

Bei den Indikatoren ‚Organisationsgrad von Gewerkschaften' und ‚zivilgesellschaftliche Dichte von Menschenrechts- und Tierschutzvereinen' fragen wir uns weiterhin, ob der unserer Meinung nach doch recht schwache und indirekte Konnex zur Demokratiequalität eine Aufnahme in den Indikatorenkatalog noch rechtfertigt. Merkel et al. argumentieren hier über einen Mechanismus der jegliches zivilgesellschaftliche Engagement als eine Stärkung demokratischer Graswurzeln begreift. Die Möglichkeit eines solchen Mechanismus erkennen wir durchaus an, allerdings besteht unserer Meinung nach hier nicht bei all den vom DB herangezogenen gesellschaftlichen Organisationen ein entsprechender Automatismus, der die Aufnahme dieses Indikators rechtfertigen würde. Wir bleiben deshalb dabei, dass Indikatoren, die wie der Gewerkschaftsgrad nur sehr indirekt mit Demokratiequalität verbunden sind und deren Ausprägungen dazu von Land zu Land, völlig unabhängig von einem Zusammenhang mit der Demokratie, aufgrund spezifischer historischer Gegebenheiten variieren, Probleme verursachen können. Gerne hätten wir in der Replik deshalb auch etwas dazu gelesen, inwiefern der Rückgang des Organisationsgrades der Gewerkschaften in der Tschechischen Republik von 80,5% 1990 auf 20,5% im Jahr 2007 wirklich dem postulierten Rückgang demokratischer Qualität zur Folge hatte (Jäckle et al. 2012, S. 111). Hier schweigen Merkel und seine Kollegen allerdings. Und auch auf die Frage, inwiefern die Einstellung zum Schwarzfahren wirklich noch sinnvoll als Maß für Demokratie gelten kann, oder ob hier nicht doch ein Fall der schrittweisen Ausweitung des Grundgedankens vorliegt, gibt die Replik von Merkel et al. leider keine Antwort.

Aber nicht nur das Überdehnen des Grundkonzepts durch nur sehr indirekt mit der Demokratiequalität verbundene Indikatoren sehen wir als Problem, gleichzeitig fehlt die nötige Schwerpunktsetzung bei deutlich relevanteren Aspekten. An dieser Stelle möchten wir das Augenmerk auf die Erfassung direktdemokratischer Institutionen lenken. Im Demokratiebarometer gibt es nur eine Variable von insgesamt 100 aufgenommenen Variablen, die die Direktdemokratie berücksichtigt. Auch diese „Geringschätzung" der Volksrechte überrascht doch sehr, gerade weil die Beteiligung und Mitwirkung der Bürger an der Entscheidungsfindung einen Wesenskern der Demokratie ausmacht. Die erhobene Variable „DIRDEM" erfasst die Existenz von fakultativen und obligatorischen Referenden sowie von

[5] Ausführlicher hierzu siehe Fußnote 2. Ein ähnliches Problem sehen wir bei der Regierungsfähigkeit, die auch in autokratischen Staaten wie Singapur oder den VAE durchaus hoch sein kann.

Hürden im politischen Prozess. Das Vorkommen eines fakultativen Referendums wird mit 0,33 auf einer Skala von 0 bis 2 bewertet. Im Gegensatz dazu weist die bereits angesprochene Frage ob Schwarzfahren entschuldbar ist (Bühlmann et al. 2013, S. 32) als einer von vier Subindikatoren des Indikators *Devbehav*, der abweichendes soziales Verhaltens bemessen soll,[6] ein höheres relatives Gewicht auf (0,25 zu 0,167) als die Existenz eines fakultativen (oder obligatorischen) Referendums. So mag zwar einleuchten, dass die öffentliche Unterstützung einer Regierung einen positiven Effekt auf die Regierungsfähigkeit hat, doch dass die persönliche Einstellung zum Schwarzfahren hier ein valider Indikator sein soll und sogar eine größere Bedeutung letztlich für die Demokratiequalität haben soll als das fakultative Referendum, ist nur schwer nachzuvollziehen.

Insofern verwundert es auch nicht, dass im langfristigen Mittel die Schweiz (1990-2005) nur Platz 14 unter 31 Demokratien einnimmt, wobei die letzten Daten für 2007 die Schweiz immerhin auf Platz 9 von 30 Demokratien einsortieren. Merkel und Bühlmann (Merkel und Bühlmann 2011, S. 36) sind hier besonders kritisch: „Die politische Beteiligung ist in der Schweiz besonders ungleich. Ein großer Teil der Schweizerinnen und Schweizer beteiligt sich nicht an der Politik. Diejenigen aber, die sich politisch beteiligen, sind vor allem die Gebildeten, Wohlhabenden, Älteren und überproportional Männer." Diese Aussage ist im Grunde falsch, da durch die Häufigkeit der Abstimmungen *(rules in use*) und der zahlreichen Möglichkeiten der direktdemokratischen Beteiligung (*rules in form*) auf allen staatlichen Ebenen eher eine stärkere Beteiligung weiterer Kreise der Bevölkerung wahrscheinlich ist. Zumindest ist die international vergleichende Sozialkapitalforschung an dieser Stelle für die Schweiz weniger skeptisch (Freitag und Stadelmann-Steffen 2011).

3. Messung

In Bezug auf die Messung wollen wir zwei Punkte aufgreifen, die Merkel et al. in ihrer Replik ebenfalls ansprechen: erstens die Frage der Integration von Expertenumfragen, zweitens Fragen der Skalierung sowie der Aggregation der Rohdaten.

3.1 Expertenbefragungen

Merkel et al. kritisieren unseren Vorschlag, die Validität des DB durch eine Anreicherung mit Indikatoren zu erhöhen, die nicht, wie dies beim DB ausschließlich der Fall ist, aus vorhandenen

[6] Die drei weiteren Subindikatoren, allesamt auf einer Skala von 1-10 gemessen, sind: Akzeptanz bzw. Entschuldbarkeit von Steuervermeidung, Betrug und dem Erschleichen von Sozialleistungen.

quantitativen Datenbanken stammen, sondern aus Expertenbefragungen gewonnen werden. Wenn sie schreiben, dass „Reliabilität eine notwendige Voraussetzung für die Validität“ (Merkel et al. 2013, S. 80) ist, stimmt dies natürlich – Reliabilität ist wichtig, das wurde und wird von uns auch an keiner Stelle bestritten. Es bedeutet jedoch keineswegs, dass die von uns geschilderte Gefahr einer Überbetonung des Reliabilitätskriteriums zu Lasten der Validität nicht vorkommen kann. Ein einfaches Beispiel hierzu: Verwendet man einen Zollstock, bei dem die Zentimetereinteilungen fälschlicherweise systematisch immer 1,5 Zentimeter weit auseinander liegen, ist es problemlos möglich, mit diesem Meterstab sehr reliabel zu messen, d.h. die wiederholte Messung mit diesem Maß würde immer zu exakt demselben Ergebnis führen. Einen zufälligen Messfehler können wir damit praktisch ausschließen, nicht jedoch einen systematischen. Dieser führt dazu, dass wir stets für einen Tisch, der real 1,5 Meter breit ist, nur eine Breite von einem Meter messen. Durch dieses *Vorbeimessen* wäre die Validität einer solchen Messung damit nicht gegeben. Wir bleiben dabei: Gerade bei so komplexen Konstrukten wie dem der Demokratie, bei dem ein einfaches direktes Messen nicht möglich ist, wodurch sich zwangsläufig das Adäquationsproblem zwischen theoretischem Konstrukt und statistischen Gattungsbegriffen noch stärker stellt als in unserem simplen Größenbeispiel, bietet ein zusätzliches Heranziehen *weicher Daten* aus Expertenbefragungen eine Möglichkeit, die Validität der Messung zu erhöhen, bzw. zumindest die Güte der über *harte quantitative Daten* gemessenen Indikatoren kreuzzuvalidieren. Die in unserer Kritik zitierten Beispiele (Messung von Parteipositionen; Sustainable Governance Indicators) zeigen, dass wir mit unserer Meinung an dieser Stelle auch nicht ganz alleine stehen. Allerdings geben wir Merkel et al. selbstverständlich Recht, wenn sie darauf hinweisen, dass eine bloße Maximierung der Anzahl befragter Experten nicht ausreicht, um *valide* Ergebnisse zu erzielen. Das haben wir auch nicht behauptet, sondern einzig darauf hingewiesen, dass über eine ausreichende Anzahl an Experten die interne Konsistenz ihrer Antworten geprüft und hierdurch die Gefahr zufälliger Messfehler reduziert werden kann, was als Reliabilitätsmaß zu werten wäre. Um valide Ergebnisse zu erzielen, muss natürlich auch bei Expertenbefragungen sichergestellt werden, dass nicht an dem interessierenden Konzept vorbeigemessen wird. Zur Frage, wie dieses Problem zu vermeiden ist, erwähnen Merkel et al. zwei wichtige Aspekte (Merkel et al. 2013, S. 80): Standardisierung der Messinstrumente und/sowie unmissverständliche Frageformulierungen. Ergänzen ließe sich noch, dass darauf zu achten ist, dass alle Experten, insbesondere im Ländervergleich, dasselbe unter den zu bestimmenden Größen verstehen sollten, also dass das, was Przeworsky und Teune (1970) mit „cross-system equivalence“ bezeichnen, gezielt im Auge behalten wird. Man sieht also, dass auch Expertenbefragungen voraussetzungsvoll sind und wohl geplant sein wollen.

Gleichwohl sind wir weiterhin der Meinung, dass die Vorteile eines kombinierten Ansatzes aus harten, quantitativen Daten und Expertenbefragungen für die Validität des Gesamtindikators überwiegen. Hierdurch entginge man auch zum Teil der Gefahr, dass unterschiedliche Quellen des DB dieselben Rohdaten verwenden und sich gegenseitig als Quelle nennen und dadurch Daten indirekt mehrfach in die Messung eingehen. Die Begründung des DB eine Vielzahl an Quellen heranzuziehen, um so systematischen Messfehlern zu begegnen (Bühlmann et al. 2013), verfängt eben nur so weit, wie die verschiedenen Quellen nicht doch letztlich alle auf dieselben grundlegenden Daten zurückgreifen und nicht *voneinander abschreiben*.[7]

3.2 Die Skalierung und Aggregation der Daten

Merkel et al. erläutern in ihrer Replik nochmals kurz die von ihnen verwendete Skalierung der Rohdaten nach dem Best-Practice-Verfahren. Leider versäumen sie es dabei, inhaltlich Stellung zu unserem Einwand zu beziehen, dass dieses Verfahren die auf konzeptioneller Ebene ausgeklammerte Bestimmung für eine Demokratie wünschenswerter Maxima und Minima bei den einzelnen Indikatoren auf ein unreflektiertes „der Wert sollte möglichst groß sein" auslagert. Ist es nicht so, dass, wie von uns angesprochen, in der Regel Länder mit Wahlpflicht eine höhere Wahlbeteiligung aufweisen, diese somit über das Blueprint Sample den Standard für gute Demokratie bei diesem Indikator setzen, wodurch letztlich einer bestimmten Demokratietheorie, nämlich der partizipativen, der Vorzug gegeben wird? Und steht dies nicht der ursprünglichen Konzeption des DB entgegen, nach der keine spezifische Austarierungsoption zwischen Freiheit, Gleichheit und Kontrolle bevorzugt werden sollte?
Merkel et al. schreiben bezogen auf die von ihnen verwendete Skalierung, dass „einer impliziten Gewichtung des Datenmaterials vorgebeugt werden" (Merkel et al. 2013, S. 81) muss, was das Best-Practice-Verfahren dadurch erreiche, dass „es die Spannweite jedes Indikators zu dessen Standardisierung heranzieht" (Merkel et al. 2013, S. 81)[8]. Wir hätten in unserer Kritik diese Problematik hingegen gar nicht erkannt. Das ist falsch. Unsere Kritik bezog sich dezidiert auf die implizite Gewichtung, die erst durch die verwendete Skalierung über die empirisch im Blueprint Sample vorgefundenen Spannweiten erfolgt. Die von Merkel et al. als

[7] Augenfällig ist dies beispielsweise bei den vom DB durchgeführten Imputationen fehlender Werte, bei denen immer wieder die Governance Indicators der Weltbank herangezogen werden, welche über Faktoranalysen selbst stets aus einer großen Masse an zum Teil auch vom DB verwendeten Indikatoren gebildet werden.
[8] Laut Merkel et al. sollen „alle Elemente auf derselben Aggregationsstufe […] gleich wichtig" (Merkel et al. 2013, S. 81) sein. Bereits die ungleiche Anzahl an Indikatoren, aus denen die 18 Komponenten aufgebaut sind, bedeutet jedoch, dass Indikatoren wie der Corruption Perception Index, der mit drei weiteren Indikatoren zusammen die Subkomponente „no secrecy" konstituiert, letztlich stärker gewichtet werden als Indikatoren wie die Wahlbeteiligung, die mit fünf anderen die entsprechende Subkomponente „equality of participation" ausmacht (vgl. Bühlmann et al. 2011).

scheinbar elegante Lösung angeführte Min-Max-Normalisierung löst das Problem der impliziten Gewichtung nicht, sondern verlagert es nur. Da die hierzu in der Replik von Merkel et al. gegebenen Erläuterungen in unseren Augen missverständlich sind, wollen wir hier das grundlegende und an sich nicht sonderlich komplexe Problem nochmals aufzeigen: Wenn die Rohdaten bei Indikator X unter den Ländern nur marginal streuen, obgleich die Skala der Rohdaten eventuell deutlich größere Streuungen ermöglichen würde, bei Indikator Y hingegen praktisch die gesamte mögliche Spannweite ausgeschöpft wird, dann sehen wir es weiterhin als problematisch an, wenn die Skalierung die marginalen Unterschiede bei X so *aufbläht*, als wären diese genauso groß wie bei Y. Hierdurch werden implizit nämlich gerade diejenigen Unterschiede und damit die betreffende Variable stärker gewichtet. Anders ausgedrückt generiert die Skalierung eventuell Varianz, wo realiter keine nennenswerten Unterschiede vorliegen.

Merkel et al. unterstellen uns eine „Fehlinterpretation" (Merkel et al. 2013, S. 81), wenn wir die ersten beiden Aggregationsschritte des DB als eine Verknüpfung hinreichender Bedingungen auffassen und ab der Ebene der Komponenten dann die einzelnen zu aggregierenden Teile als jeweils notwendige Bedingungen bezeichnen. Wir bleiben jedoch dabei. Laut Munck und Verkuilen (2002, S. 24) steht eine Aggregationsregel, die die zu aggregierenden Teile multiplikativ verknüpft (wie die Arkustangens-Formel) dafür, dass diese Elemente jeweils notwendige Bedingungen für das Aggregat darstellen. Entsprechend impliziert eine additive Verknüpfung (wie durch das arithmetische Mittel), dass die Teile jeweils als hinreichende Bedingungen für das Aggregat aufzufassen sind. Die in der Replik angeführten Erläuterungen, dass die Logik notwendiger und hinreichender Bedingungen ausschließlich zur Bestimmung des Blueprint Samples relevant sei, führen an unserem Punkt vorbei. Und auch der Hinweis dass es schwierig sei, „auf der Ebene der Indikatoren und Subkomponenten theoretische Annahmen über deren Effekt auf die Demokratiequalität zu treffen" (Merkel et al. 2013, S. 82), erklärt in unseren Augen nicht, weshalb an diesen Stellen das arithmetische Mittel als Aggregationsregel herangezogen wird, von der Funktionsebene an, ab der „ein klarer theoretischer Rahmen" (Merkel et al. 2013, S. 82) vorliegt,[9] hingegen die Arkustangensformel – womit das DB hier einen Schwenk von einer Aggregationslogik der hinreichenden zu einer der notwendigen Bedingungen vollzieht.

[9] Diese Aussage verblüfft. Gingen wir doch bislang davon aus, dass der klare theoretische Rahmen das gesamte Konzept des DB umfasst, bis hin zu den einzelnen Indikatoren.

Bezüglich der Arkustangensformel als Aggregationsregel bleiben wir weiterhin skeptisch. Sie ist unserer Meinung nach unnötig komplex,[10] was die intersubjektive Nachvollziehbarkeit unterminiert; sie erzeugt künstlich Varianz (auf diesen Einwand gehen Merkel et al. leider nicht ein); und die Anforderungen, für die diese Formel Lösungen bietet (Abbildung einer begrenzten Substituierbarkeit und eines abnehmenden Grenznutzens), muss man nicht zwangsweise aus dem ursprünglichen Konzept der drei interdependenten Grundprinzipien Freiheit, Gleichheit und Kontrolle als Ansprüche an die Aggregation herauslesen.[11]

Literatur

Bühlmann, Marc, Wolfgang Merkel, Lisa Müller, Heiko Giebler und Bernhard Wessels. 2011. *Democracy Barometer. Methodology*. Aarau: Zentrum für Demokratie.

Bühlmann, Marc, Wolfgang Merkel, Lisa Müller, Heiko Giebler, Bernhard Wessels, Daniel Bochsler, Miriam Hänni und Karima Bousbah. 2013. *Democracy Barometer. Codebook for Blueprint Dataset Version 3*. Aarau: Zentrum für Demokratie.

Bühlmann, Marc, Wolfgang Merkel, Lisa Müller, Heiko Giebler und Bernhard Weβels. 2012. Demokratiebarometer: ein neues Instrument zur Messung von Demokratiequalität. In Indizes in der Vergleichenden Politikwissenschaft. ZfVP-Sonderheft 2/2012, Hrsg. Gert Pickel und Susanne Pickel, 115-159. Wiesbaden: SpringerVS.

Freitag, Markus und Isabel Stadelmann-Steffen. 2011. Das Freiwillige Engagement in der Schweiz. Aktuelle Befunde und Entwicklungen aus dem Freiwilligen-Monitor Schweiz 2010. In *Grenzen-Los! Fokus Gemeinden. Freiwilliges Engagement in Deutschland, Österreich und der Schweiz,* Hrsg. Herbert Ammann, 104-130. Zürich: Seismo-Verlag.

Jäckle, Sebastian, Uwe Wagschal und Rafael Bauschke. 2012. Das Demokratiebarometer: „basically theory driven"? *Zeitschrift für Vergleichende Politikwissenschaft* 6 (1): 99-125.

Kaina, Viktoria. 2008. Die Messbarkeit von Demokratiequalität als ungelöstes Theorieproblem. *Politische Vierteljahresschrift* 49 (3): 518-524.

Merkel, Wolfgang und Marc Bühlmann. 2011. Die Vermessung freier Gesellschaften. Das Demokratiebarometer bietet ein differenziertes Bild. *WZB Mitteilungen* 131: 34-37.

Merkel, Wolfgang, Dag Tanneberg und Marc Bühlmann. 2013. „Den Daumen senken": Hochmut und Kritik. *Zeitschrift für Vergleichende Politikwissenschaft* 7 (1): 75-84.

Munck, Gerardo L. 2012. *Conceptualizing the Quality of Democracy: The Framing of a New Agenda for Comparative Politics*, DISC Working Paper Series. Budapest: Central European University.

[10] Aus dieser Warte betrachtet möchten wir auch unseren Vergleich mit dem arithmetischen und geometrischen Mittel verstanden wissen. Unser Anliegen war es dabei in erster Linie den Effekt, den die Arkustangensformel hat, plastisch im Vergleich zu diesen beiden bekannten Mittelwerten darzustellen. Zudem mögen die Leser für sich entscheiden, ob das geometrische Mittel wirklich so weit vom grundlegenden theoretischen Konzept entfernt ist, wie Merkel et al. uns in ihrer Replik vorwerfen, nur unter Verweis auf die Tatsache, dass es den abnehmenden Grenznutzen nicht wie der Arkustangens implementiert und die Substituierbarkeit weniger stark begrenzt. Wir halten diesen Malus im Vergleich zum Vorteil der deutlich nachvollziehbareren Aggregation für vergleichsweise gering.

[11] An dieser Stelle möchten wir auf die Abschlussarbeit von Johannes Steiniger aus Mannheim hinweisen, der eine einfache, aber in unseren Augen durchaus bedenkenswerte Alternative zur Aggregation mittels Arkustangens aufzeigt: Er schlägt vor, über die Flächen der Netzdiagramme, die das DB bereits zu Zwecken der Deskription verwendet, die Demokratiequalität zu bestimmen (Steiniger 2012, S. 52-53). Auch wenn die Anordnung der neun Funktionen die Fläche der Neunecke mit beeinflusst und man entsprechend auf Rochaden in der Anordnung kontrollieren müsste, böte dieser Ansatz ein intuitiv verständliches Maß für Demokratie, bei dem allerdings weder die begrenzte Substituierbarkeit noch der abnehmende Grenznutzen mit implementiert wäre.

Munck, Gerardo L. und Jay Verkuilen. 2002. Conceptualizing and Measuring Demoracy. Evaluating Alternative Indices. *Comparative Political Studies* 35 (1): 5-34.
Przeworsky, Adam und Henry Teune. 1970. *The Logic of Comparative Social Inquiry.* New York: Wiley.
Steiniger, Johannes. 2012. Das Demokratiebarometer: Evaluation eines neuen sozialwissenschaftlichen Instruments, Wissenschaftliche Abschlussarbeit im Fach Politikwissenschaft (10.12.2012). Universität Mannheim.