Methodological aspects in cross-national research

Hoffmeyer-Zlotnik, Jürgen H. P. (Ed.); Harkness, Janet (Ed.)

GESIS

ZUMA

NACHRICHTEN

Spezial Band 11

Methodological Aspects in Cross-National Research

Jürgen H. P. Hoffmeyer-Zlotnik,

Janet A. Harkness (Eds.)

Zentrum für Umfragen, Methoden und Analysen (ZUMA)
ZUMA ist Mitglied der Gesellschaft Sozialwissenschaftlicher Infrastruktureinrichtungen e.V. (GESIS). Die GESIS ist eine Einrichtung der *Wissenschaftsgemeinschaft Gottfried Wilhelm Leibniz* (WGL) (www.wgl.de).

Vorsitzender des Trägervereins ZUMA e.V.: Prof. Dr. Jan van Deth
Direktor: Prof. Dr. Peter Ph. Mohler

Hausanschrift	**Postanschrift**
B 2, 1	Postfach 12 21 55
68 159 Mannheim	68 072 Mannheim

Telefon	0621/1246 - 175	
Fax	0621/1246 - 100	
E-Mail	hoffmeyer-zlotnik@zuma-mannheim.de	
Internet	www.gesis.org/zuma	ZUMA
	www.gesis.org/	GESIS

ISBN 3-924220-29-8

Druck: Druck & Kopie Hanel, Mannheim

METHODOLOGICAL ASPECTS IN CROSS-NATIONAL RESEARCH: FOREWORD

Cross-national and cross-cultural survey research has been growing apace for several decades and interest in how best to do them has possibly never been greater. At the International Sociological Association Research Committee 33 (Logic and Methodology) several sessions were dedicated to cross-cultural cross-national survey methodology and the vast majority of the papers in this volume were presented at that conference.

Researchers involved in comparative research have always been worried about measurement issues, comparability, reliability and validity of their data. But the design and execution of comparative studies has changed markedly since the early cross-national projects of the nineteen sixties and nineteen seventies (Gauthier, 2000).

1 Cross-National Research

In the sixties and seventies, cross-national projects were often basically case studies. Nationally designed questionnaires on the same topic were used to collect data in different countries (see Bendix, 1963; Barnes & Kaase, 1979). Even in excluding translation in this way, measurement challenges remained at the conceptual level: "Comparative sociological studies represent an attempt to develop concepts and generalizations at a level between what is true of all societies and what is true of one society at one point in time and space." (Bendix, 1963: 532).

One of the main questions in early research was whether social phenomena observed in different social systems were comparable or not (see Przeworski & Teune, 1970: 11). Systematic errors, in this early period of cross-national research were held to be generated by

- the method(s) used to measure across countries
- differences in social and political systems studied,
- translation from one language to another language and from one culture to another culture.

"Cross-system comparisons of single variables will be dependent upon the units and the scale of measurement within each social system" (Przeworski & Teune, 1970: 42).

Overcoming language barriers was seen as one of the first steps towards comparability. Researchers learned about "functional equivalence" (but see Johnson, 1998) and about "iterative translation" processes thought to enhance face validity of questions in intercultural implementations. As a result, they either tried to improve translation efforts by checking them with "back translation" (e.g. Brislin, 1980; but see Harkness, 2003) or, giving up on translation, advocated using country-specific items selected for each population that would measure the same underlying construct across the populations in the study. Przeworski & Teune (1970) for example demanded that in intercultural research comparative indicators as well as national indicators should be applied.

Well aware of the issues of comparability, investigators used the methods current at the time to translate, to test source questionnaire questions, to collect data on socio-demographic information and to harmonise variables for merged data sets.

With direct measurement we even have problems in national survey research because national societies are different in class affiliations, in education and in cultural history (i.e. philosophy of life in different religions). The inglorious answer to the problem was often that researchers simply ignored problems of question comprehension.

2 Translation and Question Design

Working within a cognitive framework distinguishing between intended meaning of questions and perceived meaning of questions, cognitive survey methodologists have developed and refined techniques for pre-testing questionnaires (see contributions in Presser et al., 2004). In international research, survey methodologists and linguists have been refining translation techniques and strategies that are gradually replacing older arrangements for translation and testing (Harkness, 2003; Harkness et al., 2004). It has long been recommended, for example, that translation teams include expertise in surveys, survey topics, and language. Recent prominent guidelines on survey translation, such as those by the US Bureau of the Census (REF) and those for the European Social Survey (WEB), underscore this point. At the same time, the clearer perception of translation issues that has emerged points more clearly than ever to the inter-relation between question design and question translation. Since questions are perceived in context, the different contexts in which translated questions are processed (different cultures) may change the perceived meaning of a well-translated question (Braun, 2003; Harkness, 2004). Much more research is needed on how to resolve design issues better and to trace out the potential for adaptation when producing new language version of a source questionnaire.

3 The Process of Harmonisation

Demographic and socio-economic variables describe the context in which a person acts. Context variables or background variables are variables that "contain information necessary to define homogeneous subgroups, to establish causal relations between attitudes and societal facts, and to define differences between scores on scales." (Braun & Mohler, 2003: 112).

In many countries no standardised questions exist to collect data on demographic and socio-economic variables. Even in Germany, where the process of standardisation started in the late seventies, only a small number of projects use the proposed standards. In cross-national comparable research, existing standardised instruments or indices can be found for only a very small group of variables (see Hoffmeyer-Zlotnik & Wolf, 2003).

Today, international data collection programs use different techniques of harmonisation but all share a high-level of methodological consciousness.

Output harmonisation is normally *ex-post* harmonisation, that is harmonisation carried out retroactively. Output harmonisation starts from a common, internationally agreed definition for a variable representing a common indicator. The goal or the target value to be surveyed is determined. The selection of suitable survey methods is left to the participating researchers and is accomplished by a national measurement instrument using national categories. Here the national researchers should aim for the best operationalisation of the common indicator. If the measurement procedure is valid for the national as well as for the international concept, then the approach is called ex-ante output harmonization. This ideal case, where the national measurement matches international needs, is rare; national indicators usually reflect the culture for which they were developed.

The International Social Survey Programme (ISSP) carries out a blend of output and input (see below) harmonisation with regard to socio-demographic variables. As this is an ongoing programme of surveys, members know in advance what information has to be collected. The Archive guidelines on that needs to be included is quite specific and is in the process of becoming even more specific. Thus countries can tailor their national questions to collect what is required for the programme. At the same time many ISSP countries field the ISSP survey as part of a larger study and the exact formulation of the socio-demographic questions may be determined by the larger study. The ISSP has a special demographics methods work group currently working on output harmonisation issues.

Input harmonisation takes internationally agreed standards (such as definitions, concepts, aggregations, classifications) as a starting point and then uses harmonised survey methods to implement these standards. "All survey countries use precisely the same survey procedures in an ideal case. Country-specific particularities are only permissible where they are indispensable" (Information Society Technologies, 1999: 1). Input harmonisation is always *ex-ante* harmonisation.

The European Social Survey (ESS) is an example for input harmonisation. The questionnaire was developed in English language from a predominately English group of scientists. Since the income categories in the ESS, for example, are tailored to the British income structure, the categories were not suitable for some of the other participating countries. For example, in analysing the data of ESS round 1 Warner & Hoffmeyer-Zlotnik (this volume) were unable to reproduce the income structure in Portugal or Luxembourg adequately.

4 Overview of the Volume

The idea for this volume was born during the Sixth International Conference on Social Science Methodology in Amsterdam in August 2004, organised by the International Sociological Association Research Committee 33 on Logic and Methodology. Most of the contributions in this volume are proceeding papers from the Amsterdam conference.

The contributions in this volume are organised in four parts. The first part deals with designing and implementing cross-cultural surveys. *Blair & Piccinino* note that instruments developed for one country often need to be modified for use in another. Their paper discusses the types of problems encountered and points out which pre-testing techniques (including new and conventional cognitive strategies) are useful for addressing which type of problem. *Dean, Caspar, McAvinchey, Reed & Quiroz* describe new additions to the "Question Appraisal System", a coding tool for pre-testing instruments that aim to discover cross-cultural problems in source questionnaires. *Sagebiel* discusses an international comparative project on gender in engineering, outlining the potential and challenges involved in managing a complex international research design, and presenting comparisons on the issues of gender in engineering education. The contribution by *Johnson, Cho, Holbrook, O'Rourke, Warnecke & Chávez* examines the effects of questionnaire design features on cultural variations in question comprehension difficulties. *Miller, Willis, Eason, Moses & Canfield* describe their findings in relation to tested survey questions in cross-cultural research, using an analytic approach in which the coding and tabulation of results was supplemented by interviewers' open-ended text comments.

The second part consists of three papers that deal with different issues of comparability or "equivalence". *Braun & Harkness* explore the relationship between non-linguistic aspects and language-anchored features of survey questions that create problems for "equivalence" or comparability. *Rother* investigates potential problems for "functional equivalence" on the basis of the ESS immigration questions. *Zucha*'s research focuses on structural equivalence in cross-national research and uses ISSP data to examine whether the empirical construct is universal for the countries under investigation.

The third part of the volume brings together papers on with harmonising socio-demographic information in different types of surveys. *KÖRNER & MEYER* discuss approaches for harmonising socio-demographic information in comparative household surveys used in German official European statistics. *KOLSRUD & SKJÅK* present examples of input harmonisation of demographic and socio-economic variables from the European Social Survey (ESS), in which the variables were developed by a centrally co-ordinated team of experts. Finally, *SCHOLZ* presents an example of harmonisation of demographic and socio-economic variables from the International Social Survey Programme (ISSP) for educational variables based on country-specific questions and then converted to international categorical systems.

The last section of the volume contains papers that discuss individual socio-demographic variables in cross-national perspective. *WARNER & HOFFMEYER-ZLOTNIK* discuss challenges in measuring income is cross-national surveys, while *HOFFMEYER-ZLOTNIK & WARNER* focus on the measurement of education in comparative social research. Both papers include discussion of the instruments used in the European Social Survey. *GANZEBOOM* compares different approaches to occupational coding (detailed and "crude" coding). *LAMBERT* presents the different concepts of ethnicity and the different measurement instruments in social survey research, and discusses reasonable solutions from two alternative perspectives. *WOLF* takes a closer look on religious involvement and religiosity measurements in the International Social Survey Programme and the European Social Survey and reports consistent findings on the level of religiosity in European countries despite the surveys using partly different indicators.

References

Barnes, S. H., M. Kaase, et al. 1979. *Political Action: Mass Participation in Five Western Democracies.* Beverly Hills: Sage.

Bendix, R. 1963. "Concepts and Generalizations in Comparative Sociological Studies." *American Sociological Review* 28: 532-539.

Braun, M. 2003. *Funktionale Äquivalenz in interkulturell vergleichenden Umfragen. Mythos und Realität.* Habilitationsschrift.

Braun, M., and P. Ph. Mohler. 2003. "Background Variables." Pp. 99-113 in *Cross-Cultural Survey Methods*, edited by Harkness, J. A., F. J. R. van de Vijver, and P. Ph. Mohler. Hoboken NJ: Wiley.

Brislin, R. W. 1980. "The Wording and Translation of Research Instruments." Pp. 137-164 in *Field Methods in Cross-Cultural Research,* edited by Lonner, W. J., and J. W. Berry. Beverly Hills, CA: Sage.

European Social Survey (ESS). *European Social Survey, Round 2. Specification for participating countries*. http://www.europeansocialsurvey.org/

Gauthier, A. H. 2000. *The Promises of Comparative Research.* Paper prepared for the European Panel Analysis Group.

Harkness, J. A. 2003. "Questionnaire Translation." Pp. 35-56 in *Cross-Cultural Survey Methods,* edited by Harkness, J. A., F. J. R. van de Vijver, and P. Ph. Mohler. Hoboken NJ: Wiley.

Harkness, J. A. 2004. *Translation: Final Report on Round One to the European Commission. ESS.* http://europeansocialsurvey.org/

Harkness, J. A., B. Pennell, and A. Schoua-Glusberg. 2004. "Questionnaire Translation and Assessment." in *Methods for Testing and Evaluating Survey Questionnaires,* edited by Presser, S., J. Rothgeb, M. Couper, J. Lessler, E. Martin, and E. Singer. New Jersey: John Wiley and Sons.

Johnson, T. P. 1998. "Approaches to Equivalence in Cross-Cultural and Cross-National Survey Research." Pp. 1-40 in *Cross-Cultural Survey Equivalence. ZUMA-Nachrichten Spezial 3,* edited by Harkness, J. A. Mannheim: ZUMA.

Hoffmeyer-Zlotnik, J. H. P., and C. Wolf (Eds.). *Advances in Cross-National Comparison. A European Working Book for Demographic and Socio-Economic Variables.* New York: Kluwer Academic / Plenum Publishers.

Information Society Technologies and CHINTEX. 1999. *CHINTEX Synopsis. Contract no IST-1999-11101.* http://www.destatis.de/chintex/download/synopsis.pdf

International Social Survey Programme (ISSP). www.issp.org/

Presser, S., J. Rothgeb, M. Couper, J. Lessler, E. Martin, J. Martin, and E. Singer (Eds.). 2004. *Methods for Testing and Evaluating Survey Questionnaires.* New Jersey: John Wiley and Sons.

Przeworski, A., and H. Teune. 1970. *The Logic of Comparative Social Inquiry.* New York: Wiley.

JÜRGEN H.P. HOFFMEYER-ZLOTNIK &
JANET A. HARKNESS

PART I: DESIGNING AND IMPLEMENTING CROSS-CULTURAL SURVEYS

THE DEVELOPMENT AND TESTING OF INSTRUMENTS FOR CROSS-CULTURAL AND MULTI-CULTURAL SURVEYS

JOHNNY BLAIR & LINDA PICCININO

1 Introduction

There are several motivations for proposing wider and more systematic use of the instrument development and testing methods described in this chapter: the wide range of research contexts in which cross-cultural instrument development occurs suggests a methodical consideration of the design implications of these different contexts for instrument development would be useful. Measurement error research has shown that a number of factors can play an important role in response error and data quality; and the analysis of exactly how those factors function in data collection has grown more sophisticated (Biemer et al., 1991; Presser et al., 2004). The research on measurement error due to the survey instrument has developed primarily through non-cross-cultural studies and experiences, but there is no reason to believe its findings do not have important consequences for cross-cultural surveys as well. Finally, the range of pretesting techniques has grown, as has research on the strengths and weaknesses of those different techniques.

The potential importance of the research context becomes apparent when one considers the range of situations where cross-cultural or multi-cultural surveys are done. A simple listing of some of these research contexts makes it evident that the need to consider cross-cultural factors when designing a survey instrument can occur for a number of reasons. Researchers sometimes design instruments "from scratch" for use in cultures not their own. An instrument already administered in one cultural setting may need to provide comparable measurements in a different culture. Or an instrument may be designed to use across different cultures, either in a single survey or in multiple independent surveys. The crossing of cultural boundaries may or may not involve traversing national or language boundaries as well. Different language groups and quite distinct cultures are, of course, encountered within a single country, as well as internationally.

Until fairly recently, the literature on cross-cultural survey instrument development largely focused on the issue of translation. But there are many other factors that affect measurement error in addition to the accuracy of the translation. Conversational implicature affects people's interpretation of things they hear beyond the literal meaning of the words in a statement or question (Grice, 1989). The impact of question order effects may vary from one culture to another (Schwarz, 2003). Respondents' understanding of the general intent of a survey, planned uses of the data, and assurances of confidentiality are also factors that affect how they understand and respond to survey questions. Finally, some aspects of cognition that affect response behaviors and response effects can differ across cultures (Johnson et al., 1997). While there have been important developments in cross-cultural survey design (see for example Grosh & Glewwe, 2000), much of the literature and technical reports of particular surveys' methodology suggest that frequently they do not take full advantage of the available instrument development and testing techniques. Many problems encountered in designing instruments for cross-cultural and multi-cultural research certainly could be addressed using one or more of the pretesting techniques that are now in common use.

The literature on cross-cultural survey design and the technical reports of such surveys' methodologies suggest that many of the surveys do not take full advantage of the available instrument development and testing techniques. Many problems encountered in designing instruments for cross-cultural and multi-cultural research certainly could be addressed using one or more of the pretesting techniques that are now in common use.

It may well be that these techniques themselves will sometimes need to be adapted to accommodate cultural considerations, which is still another area that would benefit from careful reports of experiences on particular surveys as well as from methodological research. It is important to keep in mind that in developing cross-cultural instruments, the researcher confronts all the usual issues of writing clear questions that capture the construct of interest and that present respondents with tasks that are reasonable; but that added to these is a range of cultural and communicative issues.

Cultural issues also exist for researchers designing surveys for their own cultures and countries. However, many factors, particularly societal norms for behaviors and interactions, might be taken into account almost unconsciously. These types of issues, however, can also be unconsciously overlooked when designing a survey for another culture, transferring a survey between cultures, or designing a single survey meant to be adaptable to multiple cultures. All these diverse factors support the potential value of applying a systematic approach to defining the instrument development issues and potential problems, selecting the appropriate instrument testing methodologies and carefully implementing them.

2 Response Effects

Concern about response effects has a long history in survey research, but seems to have played much less of a role in cross-cultural survey methodology than in research and surveys that do not address cross-cultural issues. Response effects play a crucial role in survey measurement error. Much of the response effects research followed the finding of Sudman and Bradburn's meta-analysis showing that beyond the text of survey questions, the nature of the response tasks can strongly influence respondent answers (Sudman & Bradburn, 1974). In reporting the results of a large series of experiments and re-analyses Schuman & Presser (1981) provided extensive evidence of effects resulting from alternative question wording.

Up until now, little was known about the underlying pragmatic and psychological mechanisms that underlie those effects. The Cognitive Aspects of Survey Methodology (CASM) movement began to change that. A number of works have shed light on how cognitive and communicative processes influence survey response to produce some of the response effects that have been observed (Schwarz & Sudman, 1996; Schwarz, 1996; Tourangeau, Rips & Rasinski, 2000). It would seem a logical step to extend the cognitive testing and analytic approach to cross-cultural studies and certainly some important work in that direction has been done. There is strong evidence that respondents from different cultures can vary in their response behaviors in reaction to the same survey question. In an important paper, Johnson et al. (1997) make a convincing case for the potential of cultural differences to impact each stage of the response process: comprehension, recall, response formation and reporting. Such effects result, in part, from the fact that "Cultural groups are also known to vary along the dimensions of individualism versus collectivism, emotional control versus emotional expressiveness, masculinity versus femininity, and the acceptability of physical contact" (Johnson et al., 1997: 89).

Memory retrieval and judgment formation can likewise be affected by cultural differences. In particular, semantic memory – in which information storage is linked to conceptual categories – may be structured differently across cultures. Whether such differences would affect survey recall tasks is an open question. In forming some types of judgments, respondents will sometimes rely on a frame of reference (for example, what constitutes a reasonable expectation, say, for health care) or an anchor point (such as what are the norms or what constitutes "average" behavior in their culture for, say, time spent on recreation) and decide on their answer in relation to these heuristic devices. There is certainly a possibility that these frames of reference and anchor points may differ across cultures. If so, response scales may not be used in the same way, an important factor if, for example, results in multiple countries are to be compared. Finally, editing of responses may be affected by respondents' understanding of the reporting task – how much informa-

affected by respondents' understanding of the reporting task – how much information is wanted in response to an open response question or, as Johnson et al. (1997) point out, by such factors as self-presentation and some aspects of the respondent's interaction with the interviewer.

More recently, Johnson & van de Vijver (2003) reviewed what we know about culturally-impacted effects in the area of social desirability. This may be seen as a special, but illustrative, case of response editing. Their discussion focuses on whether social desirability might have a differential effect on respondents from different cultural backgrounds, which may be a function of real differences in response propensities between cultures. But on another view, social desirability may be due to characteristics of the survey question, the interaction between the interviewer and respondent, or both. In particular, they point out that "cultural differences between respondents and interviewers sometimes produces varying patterns of responses." Additionally, they note that survey administration mode can affect reporting, particularly on sensitive questions.

In summary, for many factors affecting response behaviors there is some theoretical basis for expecting cultural differences, but for most, only a small amount of survey-based research has been done. But even the little we know at this point suggests caution in assuming how well survey instruments will "travel" across cultures; and to use multiple 'tools' to assess how an instrument will perform.

Although this research has not focused on the implications of these findings for pretesting instruments that have been created or transformed to work in different cultures, clearly it is a significant issue. As far as can be determined from the literature much cross-cultural instrument development and testing has been concerned with translation. Beyond issues of the quality of the original question's translation, as has been noted, comprehension can be different or difficult (not the same issues) due to cultural factors. Moreover, it is often unclear how much pretesting was done on an instrument before using it in the researcher's *home culture*, let alone how its performance might be affected by transplanting it into another cultural context. It would seem reasonable, considering the research noted above and even an informal observation of common practice in cross-cultural surveys, that those surveys might benefit from using a wider array of pretesting methods. How that might best be done and how those methods themselves need to be adapted to the cross-cultural research context is a subject for future methodological research and practical utilization.

3 Review of Some Issues in Instrument Testing

3.1 Comparability of scales

Another important component of the pretesting and testing process is to uncover societal similarities and differences in the measurement of frequencies of unobservable and observable behaviors. One way to accomplish this is to design and test a set of response scales that are thought to be comparable from one cultural context to another. Depending on the type of response scale used, conclusions drawn about differences in behaviors and their frequency of occurrence can be false or misleading. It is only when the items in questionnaires representing multiple languages and contexts have equivalent response scales that comparable measurement and valid inferences can be obtained (Smith & Wolter, 2004).[1]

Ji, Schwarz & Nisbett (2000) used a set of open-ended responses and a set of frequency scales to explore cultural differences in behavior and memory. They found that the when the frequency scales were used, these had differential effects on reports of frequency of observable behaviors and unobservable behaviors in China and the United States. Ji et al. (2000) also reported that when the open-ended response format was used, it yielded results for observable behaviors that were about equal for China and the United States (as was intended, by design). They concluded that, depending on the response format used, researchers could arrive at different conclusions about cultural differences in behavior. This study illustrated the potential risks associated with the use of frequency scales in cross-cultural research and underscored the need for pretesting of response scales.

3.2 Equivalence

The question of equivalent measures is often an issue in cross-cultural surveys. Such equivalence becomes a central concern when a single survey encompasses more than one cultural setting, or when the same survey is conducted for the purpose of comparing different cultures or countries. There are essential multiple dimensions of equivalence of measures: equivalent comprehension of questions, including response scales, and equivalent respondent use of response scales (as discussed above).

These dimensions seem central, though there is by no means agreement on this point.

Johnson (1998) provides a comprehensive review of alternative concepts of equivalence, listing fifty-two types of equivalence identified in a literature review; and goes on to

1 In their paper, Smith & Wolter (2004) offer some preliminary ideas about what kinds of scales might lead to more equivalent cross-cultural comparisons.

consider their impact on the focus of survey design. Many of these types differ only slightly from one another; and the types also range from high-level, conceptual notions such as *complete equivalence* and *cultural equivalence* to what appear to be very literal notions of equivalence such as *text equivalence*. Among these is *measurement equivalence*, sometimes defined as "...instance in which factor loadings and error variances are identical across groups" and *semantic equivalence* which occurs when "...survey items ...exhibit identical meaning across two or more cultures after translation." These last two categories, the former of which would seem to encompass the question's response categories, seem to suggest a reasonable direction toward an operational definition of equivalence. Additionally, this direction would seem also to lend itself to developing practical procedures for attaining equivalence. We make some suggestions for beginning to specify such procedures.

The first step is the identification of points of non-equivalence. Expert review, focus groups and cognitive interviews can all be potentially useful in such identification. But, unless the degree of non-equivalence is large, the amount of effort and sample sizes necessary to accomplish this identification may be considerably greater than that need for typical pretesting to determine comprehension and other response task performance. A sequential process approach may be useful in this stage. For example, the in-country expert or an expert panel may be able to identify areas of possible non-equivalence. The principal investigator or data analyst would then judge the possible implications of non-equivalence, considering the variables affected, the rough magnitude of the problem and the likely proportion of the sample that may be impacted. On the basis of this assessment, a decision would need to be made about the importance of removing or adjusting for the non-equivalence. Just as with other sources of measurement error, it is seldom possible to correct every potential flaw. This process of identification and assessment will become more reliable over time as experience is gained with different populations and survey topics and measures.

It may be possible to revise the questions or other features of the instrument to achieve equivalence. Failing that, it may be possible to determine the relationship between non-equivalence measures and allow for it in data analysis. For example, if in a survey that includes young Hispanic males it is determined both that they overstate, say, their health status in comparison to young men from other ethnic groups, and some estimate of the extent of the overstatement, a statistical adjustment may be possible. Although the process we suggest may be costly and relatively complex, these factors have to be weighed against the potential impact of non-equivalence on key analyses.

3.3 Validity

The importance of construct validity is to ensure that the items in the instrument designed for another cultural context capture what you are attempting to measure, and are valid representations of what you are trying to measure, within that country or culture.

Miller et al. (1981) in their early paper asked what comprised a reliable index of a concept and its meaning for a specific country, and assessed the parts of the index that were valid across countries (in this case, the United States and Poland). They began their analysis by examining the *within-country validity* of the items in the US survey that were to be compared to the Polish replication survey. The analysis looked for internal consistency of the items supposed to measure the key concept (i.e. authoritarian-conservatism): empirical differentiation of the items in the key concept from other related concepts (i.e. personality), any covariation due to measurement error, and qualities of the index that showed it was an adequate representation of the concept.

They next examined the *cross-national validity* of the items in the index of authoritarian-conservatism by looking at the correlation of the within-country index of one country against the index derived from factor analysis of the other country's data. They looked at the statistical properties of indicators that were common to both countries and of those that were country-specific.

The investigators concluded that there also was a set of core items that had an equivalent meaning in both countries, but that there were country-specific items, and that those items should remain specific only to that respective country's index.

Some researchers (Miller et al., 1981) recognized the utility of certain statistical tools (e.g. factor analysis) for assessing construct validity in cross-cultural survey instruments. Others such as Saris, van der Veld & Gallhofer (2004) pointed to researchers like Campbell & Fiske (1959) who maintained that validity could best be evaluated if more than one method was used to measure the same trait; in this way errors could be detected. The methods they refer to could be as simple as using multiple versions of a response scale and comparing their correlations. Validity could then estimated from the strength of the relationship between the trait of interest and the "true score" (defined as the component of the observed variable that represents the trait and method used).

4 Pretest Techniques

The available pretesting techniques are well known. The potential strengths and shortcomings of these techniques have been noted in a number of sources based on experimental studies (Presser & Blair, 1994; Presser et al., 2004) or books that provided practical advice for implementation (Fowler, 1995; Czaja & Blair, 2005), proposed process quality frameworks for pretesting (Blair & Piccinino, 2004), or considered pretesting techniques from a theoretical perspective (Martin, 2001). All of these perspectives are potentially important for the application of the techniques in cross-cultural survey instrument pretesting. Below we list the main techniques and note selected theoretical and practical issues that may be relevant to cross-cultural surveys, as well as some suggestions about how techniques might be adapted to address some of the issues we have raised.

Presser & Blair (1994) compared four pretesting methods using a single questionnaire in repeated trials of each method. The methods tested were: conventional pretests, behavior coding, cognitive interviews, and expert panels. A model-based coding scheme was used that classified problems as respondent-semantic, respondent-task, interviewer-task, or analysis. On average, expert panels were found to be the most productive measured by the total number of problems identified. Expert panels and behavior coding were more reliable than other methods in the number of problems identified across trials, as well as in their distribution of problems.

The findings that are probably most important in their implications for cross-cultural studies concern respondent-semantic and respondent-task problems. Cognitive interviews were consistently better at identifying respondent-semantic problems; while conventional pretests and expert panels were best at identifying respondent-task difficulties.

Below we provide an overview of the main pretest techniques and note additional theoretical and practical issues that may be relevant to cross-cultural surveys, as well as some suggestions about how techniques might be adapted to address some of the issues we have raised.

4.1 Expert review

Expert reviews are a generic term for a number of different activities, involving different kinds of experts whose advice is elicited in different ways. Most relevant here are experts about the particular country or culture where the survey will be conducted; and experts in the language, dialect or patois in which the survey will be conducted. What makes a person an expert? On some level, just being indigenous or fluent in a language may qualify as sufficient expertise. It also is important to be aware of possible within-culture class

differences. Certainly different experts may come to different conclusions and provide conflicting advice or solutions to question problems. Even conflicts of this type may be useful in identifying issues that need additional attention in design or testing.

The problem of selecting experts may be true for any survey, but is particularly so in other cultures, where the principal investigator or survey methodologist might not be competent to identify "experts." One approach for helping with this problem that can have broader value as well is the use of expert panels. In this case, usually three or more experts are brought together for a discussion of pertinent survey issues. The experts need not all have the same specialty to be informative about a particular issue. One expert may have a good overall understanding of, for example, how the health care system functions, but be unfamiliar with access difficulties that certain subpopulations in the country experience. While another expert may understand how people living in cities normally handle certain financial matters, but know little about how the same functions work in poor, rural areas.

For example, in a Willingness to Pay survey in Kenya (McGunnigle et al., 2000), an examination of the data showed that respondents and interviewers had problems with survey questions about bid price (price willing to pay), and often entered numeric amounts to yes/no questions. Also, service fee categories in the instrument often did not match actual services being offered. Some discussion with the staff in Kenya after-the-fact was necessary to understand how the service and fees systems worked, and to realize problems with the data. These problems might have been minimized if discussions with Kenya experts occurred before the final development stage of the questionnaire.

When recruiting an expert panel, it is useful to both cover the range of areas considered important for the survey at hand, but also to have some overlaps in expertise so that alternative judgments or points of view can be identified and assessed.

4.2 Cognitive interviewing

The standard four-stage response model (e.g. Tourangeau & Rasinski, 1988) posits a sequential set of cognitive tasks respondents must perform in answering survey questions. This descriptive model has provided a useful general framework for instrument testing as well as other aspects of data collection.

Cognitive interviewing is a generic term for a set of available techniques that can be used in various combinations in one-on-one pretest interviews (Conrad & Blair, 2004). In large part, cognitive interviews include respondents thinking aloud, reporting everything that comes to mind as they answer the questions. Thinking aloud is combined with two types of probes: concurrent, asked during the interview; and retrospective, asked after the inter-

view (or after self-contained interview sections). The probes may be partly written in advance of the cognitive interview as well as improvised during the interview. Cognitive interviewers probe based either on respondents' indications of difficulty (e.g. requests for clarification, changing answers, inability to produce an answer) or on conjectures about aspects of questions that may cause response problems. There is evidence that the former type of probe is more reliable and less likely to turn up false positives (i.e. "identify" nonexistent problems) than the latter type, though these context-free probes may produce more problems as well (Conrad & Blair, 2004).

There is not agreement, however, on the practical implications of these findings. Willis (2004) in "Cognitive Interviewing: A Tool for Improving Questionnaire Design," the most comprehensive treatment of the method to date, takes a more positive view of using what he calls *proactive* probing. Until some methodological research on the technique has been conducted in a cross-cultural setting, caution is advised in the selection of probes.

The main strength of cognitive interviewing is its potential ability to identify problems at any stage of the response process: comprehension, recall, response formation or response reporting. Cognitive interviews often uncover possible reasons for the occurrence of identified problems; such information can help guide question repair. In addition to basic response tasks problems, cognitive interviewing can probably help identify pragmatic communicative issues that may be especially important in cross-cultural surveys. But there is only a small amount of research to date to support this contention.

It is useful to keep in mind that the tasks required in some variants of cognitive interviewing may be very difficult for some respondents. These are not things that people commonly are requested to do; some detailed explanation or examples may be required. Some cognitive interview tasks, like thinking aloud or paraphrasing, require first that the respondents understand what they are being asked to do.

Even if respondents do understand the task, those who may be less articulate or less socially comfortable talking in front of a stranger may have difficulties with some forms of cognitive interviewing. If there are possible cultural barriers to the type of interaction necessary for the successful conduct of cognitive interviews, the in-country experts may be able to point this out early in the planning process. In such a situation, the researcher may choose a version of cognitive interviewing that takes account of such barriers or decide not to use cognitive interviewing at all. Of course, this requires that the researchers clearly explain the cognitive interview process to the in-country experts.

Careful thought should be given to these issues, since difficulty with the cognitive response task may sometimes be mistaken for difficulty with the survey response task.

Looked at from a different perspective, Groves et al. (1992), in discussing direct questioning of respondents as a method to identify problems with meaning, note that "...the questions on meaning would themselves be subject to large measurement errors."

However, if target population members can, on the whole, express themselves adequately, and are comfortable responding to probes and completing other necessary cognitive interview tasks, cognitive testing can be an efficient method of uncovering a range of problems and possible solutions as well.

In work associated with a project in the Philippines in 2003, both structured and self-administered questionnaires for various types of family planning/health care providers were pretested. Through cognitive testing it was discovered that providers (especially midwives) found the self-administered portion of the questionnaire to be irrelevant and therefore did not pay much attention to answering it (Commercial Market Strategies project, 2003).

4.3 Focus groups

Focus groups can take different forms for different purposes. A focus group can consist of experts (as described above) who provide insights into the target population's country, culture, and relevant aspects of their language, or comment on any aspect of the survey. Typically, however, focus groups are composed of target population members. Just as in U.S. focus groups, decisions need to be made about what group composition will best foster the open exchange necessary to produce useful information – whether about particular subject matter or reactions to actual draft survey questions. The project's in-country experts may judge whether or not focus group interactions can be expected to work as necessary, or if not, what sorts of adaptations may be possible.

For example, in a reproductive health study in Jamaica (Young, 2003), focus groups and role-play testing prior to final questionnaire development revealed differences in attitudes of pharmacists toward youth depending on the gender of the youngster and whether they were in or from an inner-city or suburban area. Mystery client scripts were modified to incorporate the separate pharmacy needs of girls and boys, as well as to accommodate the "uptown" and "downtown" language or slang used by these youth. Even though the original instrument developed in the United States was in English, the cultural and language differences in English usage were substantial enough to warrant testing.

In another study, an evaluation of a survey in Kenya (McGunnigle et al., 2000) suggested that if the project had a longer time frame, focus groups could have provided more systematic information on the fee structure of clinics than relying solely on information from client exit interviews.

Still another type of problem was encountered in a survey in Albania (Partners for Health Reform*plus*, 2004a), and to some extent in Kenya, where little variability was found in the five-category response options to questions about client satisfaction with health providers and services. Clients generally reported satisfaction with providers/services as good or excellent, but when asked informally, some admitted they were giving "polite" answers in the structured interviews and had actually experienced a lesser degree of satisfaction than reported. Focus groups and consultation with in-country experts about customs, politics and etiquette in the region might have revealed that it was customary not to voice negative opinions publicly about health care providers/staff. This might have been remedied by permitting the survey to be modified accordingly, perhaps using a different data collection mode, and using a more elaborate introduction to try to alleviate the tendency for the respondent to use non-negative responses.

4.4 Conventional pretesting

Conventional pretesting is so named because it is the most common form of pretesting and, absent other description, what one would assume if told only that a pretest had been conducted. The technique is based on a kind of emulation of the survey. A small sample of respondents from the target population is sampled and the survey is administered to them just as intended in the actual study. A structured debriefing is held afterward in which the interviewers give their overall and their question-by-question assessment of how the interview went, what problems respondents experienced and, possibly, what changes might improve the instrument.

While the conventional pretest may include a post-interview debriefing to supplement the interviewers' impressions, usually the interviewers simply serve as proxy reporters for problems respondents had. Behavior coding (described below) can also be incorporated into conventional pretesting. Both the post-interview respondent debriefing and behavior coding can serve to validate (or not) some of the interviewers' reports.

In cross-cultural surveys, particularly if an in-country contractor is used, it is important that the interviewers go through a pretest training that not only covers issues planned for the full project interviewer training, but also discusses (with examples) the kinds of information that they are expected to be able to report about in the debriefing. They should be encouraged to take notes either during or immediately after each interview to use in the debriefing.

In a recent project in Rwanda, two days of adequately planned interviewer pretest training was truncated to two hours due to unavoidable administrative and managerial resource cuts. As a result, interviewers performed poorly in the field and had to be subjected to retraining mid-way through the field process (Partners for Health Reform*plus* (2004b).

4.5 Respondent debriefings

As noted, conventional pretests can be supplemented in different ways to obtain more and richer information about how the instrument and specific items performed. Immediately after the interview, respondents can be asked by the interviewer what they thought particular questions meant, what items they thought were difficult and why, among other things. Of course, the list of potential problems developed by the pretest team should inform the debriefing interview questions. If cognitive interviewing precedes conventional pretesting, often issues identified in the cognitive testing can inform the construction of the respondent post-interview debriefing.

The debriefing questions may be in either an open- or closed-response format. But there is some evidence (Groves et al., 1992) that open- and closed-response debriefing items may produce different information and, more importantly, present different types of response issues for the respondents. They note that respondents who are more articulate may mention issues not noted by others. But this might give a false sense of the likelihood of such problems occurring and of their distribution across the population. Two lessons that might be taken from this result: first, a mix of types of questions may be better than relying on just one kind; second, one should bear in mind the essentially qualitative nature of pretesting (and the typically small samples), and not expect to learn too much about problem distributions that might occur if flaws were left uncorrected.

4.6 Behavior coding

Behavior coding is based on a conception of what the question-and-answer interview process should be like in the absence of question flaws (Fowler, 1995). Ideally, the interviewer will read the question verbatim, without error, and the respondent will select one of the offered response options. When the process fails, certain kinds of behaviors are likely to be seen, such as respondent interruptions and requests for re-reading or for clarification, interviewer reading errors and the like. These deviations from the ideal interviewer-respondent interaction are taken as indicators of question problems if these occur relatively frequently; with problems occurring in 15% of question administrations a commonly-used threshold.

This technique does raise some questions. Should we expect the same set of 'indicator' behaviors in cross-cultural surveys? Are there supplemental codes that can capture behaviors specific to testing in other cultures? Will people in other cultures indicate problems in the same way that respondents in the United States do? Will they volunteer comments, ask questions, ask for repeats etc., if given license to do so? This is one sort of thing to check with the in-country experts – will the assumptions of this or other testing methods work as

expected in the other cultures? Ultimately, these kinds of questions need to be addressed through a combination of careful methodological research and ongoing technical reports of actual survey experiences.

5 Overview of Process Approach to Pretesting

A process quality approach to questionnaire testing is a way of ensuring coverage of the range of potential types of problems and relevant issues that can occur in different 'realizations' or contexts of cross-cultural surveys. The idea is that after specifying the type of cross-cultural instrument/survey situation, one considers the particular set of potential problems that might arise. This set of possible problem areas guides the selection of pretest techniques to address these areas.

Blair & Piccinino (2004) offer an approach for systematizing the stages involved in cross-cultural instrument pretesting. One intention of the process approach they propose is to ensure a thorough coverage of the tasks, issues and potential problems associated with instrument development. These tasks, issues and problems can be, for example, testing parameters, cultural concerns, and specific defects in questions and supporting materials. This approach requires the collaborative effort of team experts to ensure that proper coverage occurs. Survey instruments often are flawed for reasons that, in retrospect, appear quite simple and apparent. The emphasis on potential problem coverage is recognition of this. It is important to be sure "all the rocks have been turned over" in the search for problems.

Obviously a translated instrument may perform differently in a new language when exactly equivalent words are not available; or, more seriously, when no comparable concept exists in the target culture. Less obvious is that an instrument might contribute to measurement error due to the tasks required of the respondent being at odds with either respondent capabilities or when survey questions are inadvertently in conflict with some cultural norms or expectations.

As an example, questionnaires used in a survey of family planning providers in the Philippines in 2003 (Commercial Market Strategies, 2003) were implemented in English and also in Tagalog. Although the translation into Tagalog was not thoroughly tested, pretesting of the questionnaire helped reveal that respondents found the dialect of Tagalog used in the translation to be too literal so that they sometimes missed the true meaning or interpretation of the terms used. The pretesting also indicated that a more conversational version of Tagalog was preferred by respondents.

Approaches to translation involving various team design and review procedures can address many of these issues (Harkness, 2003; Harkness et al., 2002; de la Puente, Pan & Rose, 2003). The focus of a process quality approach is broader than translation, and applies even when translation is (strictly speaking) not necessary. The approach should encompass the pragmatics of communication, and even extend to practical data collection implementation issues. The process approach recognizes that both technical design issues and operational issues – such as interviewer behaviors, or obtaining thorough reports of pretest results from an in-country contractor – could affect measurement error. The strength of a process approach is the use of a general framework that can be adapted to specific surveys. As the approach is used in different surveys it is likely that variations on that general framework will develop. The dissemination of the documentation of such variations will be essential to the continued development of this approach.

6 Summary

In the above discussion, we have indicated instrument development and testing issues for which alternative combinations of pretesting techniques may prove useful. In addition, we have suggested potential problems one may encounter and possible adaptations of the techniques to make them more suitable for cross-cultural instrument testing. It is important to note that these recommendations are based on judgment and experience, rather than on methodological research. Clearly, it will be necessary to conduct such research to learn how to use these methods to best effect. Even after such research results begin to become available, it still will be necessary to consider each survey's particular characteristics, subject matter, mode of administration and target population.

References

Biemer, P. P., R. M. Groves, L. E. Lyberg, N. A. Mathiowetz, and S. Sudman. 1991. *Measurement Errors in Surveys.* New York: John Wiley & Sons.

Blair, J., and L. Piccinino. 2004. Paper presented at the RC33 Sixth International Conference on Social Science Methodology in Amsterdam, The Netherlands, August 16-20, 2004. (Also published in proceedings volume, ZUMA, May/June 2005.) *A Process Quality Approach to the Development and Testing of Cross-Cultural Survey Instruments.*

Campbell, D., and D. Fiske. 1959. "Convergent and Discriminant Validation by the Multitrait Multimethod Matrices." *Psychological Bulletin* 56:81-105.

Commercial Market Strategies (CMS) project. 2003. *Personal communication.*

Conrad, F., and J. Blair. 2004. "Data Quality in Cognitive Interviews: The Case of Verbal Reports." Pp. 67-87 in *Questionnaire Development, Evaluation and Testing Methods,*

edited by Presser, S., J. Rothgeb, M. Couper, J. Lessler, E. Martin, J. Martin, and E. Singer. New York: John Wiley & Sons.

Czaja, R., and J. Blair. 2005. *Designing Surveys: A Guide to Decisions and Procedures*. Thousand Oaks: Pine Forge Press.

Fowler, F. J. Jr. 1995. *Improving Survey Questions: Design and Evaluation.* Thousand Oaks: Sage Publications.

Grice, P. 1989. *Studies in the Way of Words.* Cambridge: Harvard University Press.

Grosh, M., and P. Glewwe (Eds.). 2000. *Designing Household Survey Questionnaires for Developing Countries. Lessons from 15 Years of the Living Standards Measurement Study.* Washington: The World Bank.

Groves, R. M., N. H. Fultz, and E. Martin. 1992. "Direct Questioning about Comprehension in a Survey Setting." Pp. 49-61 in *Questions about Questions: Inquiries into the Cognitive Bases of Surveys,* edited by Tanur, J. M. New York: Russell Sage Foundation.

Harkness, J. 2003. "Questionnaire Translation." Pp. 35-56 in *Cross-cultural survey methods*, edited by Harkness, J., F. J. R. van de Vijver, and P. P. Mohler. Hoboken: John Wiley & Sons.

Harkness, J., A. Schoua-Glusberg, and B. Pennell. 2002. *Questionnaire Translation and Questionnaire Design.* Proceedings of the International Conference on Questionnaire Development, Evaluation, and Testing Methods (QDET), Charleston, South Carolina, November 14-17.

Johnson, T. P. 1998. "Approaches to Equivalence in Cross-Cultural and Cross-National Survey Research." In *Cross-Cultural Survey Equivalence,* edited by Harkness, J. Mannheim, Germany: ZUMA News Special, Vol. 3.

Johnson, T. P., D. O'Rourke, N. Chavez, S. Sudman, R. Warneke, and L. Lacey. 1997. "Social Cognition and Responses to Survey Questions among Culturally Diverse Populations." Pp. 87-113 in *Survey Measurement and Process Quality,* edited by Lyberg, C., P. Biemer, M. Collin, C. Dippo, E. deLeeuw, N. Schwartz, and D. Trewin. New York: John Wiley & Sons.

Johnson, T. P. and F. J. R. van de Vijver. 2003. "Social Desirability in Cross-cultural Research." Pp. 195-204 in *Cross-Cultural Survey Methods*, edited by Harkness, J., F. J. R. van de Vijver, and P. P. Mohler. Hoboken: John Wiley & Sons.

Ji, L., N. Schwarz, and R. E. Nisbett. 2000. "Culture, Autobiographical Memory, and Behavioral Frequency Reports: Measurement Issues in Cross-Cultural Studies." *Personality and Social Psychology Bulletin* (May 2000).

Martin, E. 2001. *Theoretical Perspectives on the Question and Answer Process: Implications for Pretesting.* Proceedings Quest 2001, U.S. Census Bureau, Washington, DC, October 24-25, 2001, pp. 6-19.

McGunnigle, M., L. Piccinino, and T. J. Ryan. 2000. *Kenya Survey on Willingness to Pay and Client Satisfaction.* Bethesda, MD: Pathfinder/Kenya, Abt Associates Inc.

Miller, J., K. M. Slomczynski, and R. J. Schoenberg. 1981. „Assessing Comparability of Measurement in Cross-National Research: Authoritarian-Conservatism in Different Sociocultural Settings." *Social Psychology Quarterly,* 44 (3):178-191.

Partners for Health Reform*plus*. 2004a. *Primary Health Care Reform in Albania: Baseline Survey of Basic Health Service Utilization, Expenditures and Quality.* Bethesda, MD: The Partners for Health Reform*plus* Project, Abt Associates Inc.

Partners for Health Reform*plus*. 2004b. *USAID/Government of Rwanda National Health Accounts project.* Personal communication.

Presser, S., M. Couper, J. Lessler, E. Martin, J. Martin, J. Rothgeb, and E. Singer (Eds). 2004. *Methods for Testing and Evaluating Survey Questionnaires.* Hoboken: John Wiley & Sons.

Presser, S., and J. Blair. 1994. "Survey Pretesting: Do Different Methods Produce Different Results?" Pp. 75-104 in *Sociological Methodology, Vol. 24,* edited by Marsden, P. Washington: American Sociological Association.

de la Puente, M., Y. Pan, and D. Rose. 2003. Paper presented to the Federal Committee on Statistical Methodology Research Conference, Arlington VA, November 17-19, 2003. *An Overview of Proposed Census Bureau Guidelines for the Translation of Data Collection Instruments and Supporting Materials.*

Saris, W. E., W. van der Veld, and I. Gallhofer. 2004. "Development and Improvement of Questionnaires Using Predictions of Reliability and Validity." Pp. 275-297 in *Methods for Testing and Evaluating Survey Questionnaires,* edited by Presser, S., M. Couper, J. Lessler, E. Martin, J. Martin, J. Rothgeb, and E. Singer. Hoboken: John Wiley & Sons.

Schuman, H., and S. Presser. 1981. *Questions and Answers in Attitude Surveys. Experiments on Question Form, Wording and Context.* Orlando: Academic Press.

Schwarz, N. 1996. *Cognition and Communication: Judgmental Biases, Research Methods and the Logic of Conversation.* Mahwah : Lawrence Erlbaum & Associates.

Schwarz, N. 2003. "Culture-Sensitive Context Effects: A Challenge for Cross-Cultural Surveys." Pp. 93-100 in *Cross-Cultural Survey Methods,* edited by Harkness, J., F. J. R. van de Vijver, and P. Ph. Mohler. Hoboken: John Wiley & Sons.

Schwarz, N., and S. Sudman. 1996. *Answering Questions: Methodology for Determining Cognitive and Communicative Processes in Survey Research.* San Francisco: Jossey-Bass.

Smith, T. W., and K. M. Wolter. 2004. Paper presented to the American Statistical Association, Toronto, August 2004. *Techniques for Calibrating Response Scales across Countries and Languages.*

Sudman, S. and N. Bradburn. 1974. *Response Effects in Surveys: A Review and Synthesis.* Chicago: Aldine.

Tourangeau, R., and K. A. Rasinski. 1988. "Cognitive Processes underlying Context Effects in Attitude Measurement." *Psychological Bulletin* 103:299-314.

Tourangeau, R., L. Rips, and K. Rasinski. 2000. *The Psychology of Survey Response.* Cambridge: Cambridge University Press.

Willis, G. 2004. *Cognitive Interviewing: A Tool for Improving Questionnaire Design.* Newbury Park: Sage Publications.

Young, B. 2003. *Evaluating the Knowledge and Attitudes of Health Providers in the Provision of Emergency Contraception Pills and Condoms to Jamaican Adolescents.* Commercial Market Strategies (CMS) project. Kingston: JA Young Research Ltd.

DEVELOPING A LOW-COST TECHNIQUE FOR PARALLEL CROSS-CULTURAL INSTRUMENT DEVELOPMENT:

The Question Appraisal System (QAS-04)

ELIZABETH DEAN, RACHEL CASPAR, GEORGINA MCAVINCHEY, LETICIA REED & ROSANNA QUIROZ

Many approaches are used to prepare instruments for multicultural administration, depending on the scope, schedule and budget of the study. Sequential questionnaire development, the most common approach to developing cross-cultural instruments, is also the most affordable. Designers formulate and pretest an instrument in the source language, then translate it into the target language(s) using culture-specific tailoring. In contrast, parallel development incorporates target cultures throughout the design and pretesting process. The disadvantages to parallel development are that it is expensive, time-consuming, and subject to version control problems. The Question Appraisal System (QAS) is a coding tool for pretesting instruments (Willis & Lessler, 1999). The QAS is supported by an item taxonomy of the cognitive demands of a question and documents the features that may lead to response error. Results of the appraisal are used to revise question wording, response wording, questionnaire format, and question ordering. This paper describes research conducted to update the QAS to identify problems due to cross-cultural and cross-linguistic application of questions.

1 Introduction

Globalization tendencies and demographic shifts demand understanding between cultures as well as the technical ability to communicate effectively in multiple languages (Cronin, 2003). The need for understanding other cultures and languages is such that institutions of higher education now require diversity training in most fields of study. Within the United States, demographic changes in minority populations have brought about cultural diversity that is often accompanied by linguistic gaps. Latin America and Asia are the points of

origin for over 75 percent of the foreign born U.S. population, and meeting linguistic demands is a challenge for health and social service agencies, educators, policy planners and researchers. As the needs for information from diverse populations rise, survey research is charged with finding ways to bridge linguistic and cultural gaps and ensuring accurate representation in research studies (Flores et al., 2002; Li et al., 2001).

A compelling need for sound methodological practices in instrument translation and adaptation has been present in survey research for over half a century. Early studies on cross-cultural congruence revealed that wording and translation were considered 'the weakest link' in the process of attaining comparable tools for research purposes (Kumata & Schramm, 1956). During the late 1960s and early 1970s, multinational survey projects appeared, and most cross-cultural researchers dealing with human behaviour favoured investigations across the globe. Benefits derived from cross-cultural research, such as increasing the range of analytic variables and increasing sensitivity to context, would have been impossible to attain in mono-cultural research (Brislin, 1993). The needs of cross-cultural survey research, coupled with the desire to reap its benefits, have pushed the field towards establishing sound practices in instrument adaptation and translation.

In the U.S., standard practices for conducting Spanish interviews have evolved along with our understanding of cross-cultural methodology. Attending to cultural and linguistic differences among populations when developing a new instrument is costly and time consuming, therefore, using existing questionnaires of mono-cultural context is the most frequently used approach in questionnaire development. Although relatively inexpensive, the use of an existing instrument often fails to address semantic, conceptual and normative equivalence. Neglecting to address the comparability of questions can decrease the validity and reliability of measures. Finding sound but inexpensive methods of addressing cultural and linguistic issues during instrument design is a goal that survey research has not yet met. The many approaches developed so far require trade offs between cost (both in terms of money and time) and instrument quality.

A possible solution to incorporating language awareness into the developing stage of an instrument at a relatively small cost is to address potential linguistic and cultural issues in the pre-testing stage. The Question Appraisal System, or QAS-99, (Willis & Lessler, 1999), a coding system for identifying question characteristics likely to result in response errors, is capable of accommodating steps that assess potential language and cross-cultural problems. This research examines a variety of problems that cross-cultural survey designers have encountered throughout the years and proposes an enhanced version of the QAS-99, the QAS-04, as a practical and relatively inexpensive way to improve methodological practices in the cross-cultural research field.

2 Theoretical Framework

To fully understand the implications of developing survey instruments for translation and application across multiple cultures, it is important to place these activities in the wider context of communication between cultures. Too often, survey researchers develop measures without understanding how behaviours are understood differently across cultures. To understand the context of a survey response, researchers must know how concepts, values, and linguistic constructions vary across languages and cultures. Furthermore, researchers must understand how cultures vary simply in their demographic characteristics. This section provides a brief summary of these cross-cultural variations.

2.1 Understanding the link between culture and response

A hypothesis or research question often originates within a specific cultural context. What may be an important question in one culture may not be important or even relevant to people of a different culture. Only when the population studied understands, is able, and is willing to answer a question should other methodological issues be addressed (Fowler, 1995; Peterson, 2000). In adapting an instrument, researchers must aim to establish comparability of concepts, norms and semantics, none of which can be achieved without understanding the culture of the respondents.

Concept comparability. Brislin (1993) notes that understanding behaviour within context can provide insights on responses that otherwise might be attributed to the wrong cause. To establish a common understanding of concepts when studying different cultures, researchers have called the culture-common and culture-specific perspectives 'emic' and 'etic'. The emic perspective involves the evaluation of a studied phenomenon from within the culture and its context and it aims to understand its significance and its interrelations with cultural elements. The etic perspective, in contrast, involves the evaluation of phenomena from outside the culture, aiming to identify and compare similar phenomena across different cultures (Berry, 1969; Brislin, 1993).

Failing to become acquainted with emics of populations being targeted in survey research can lead to misattribution of response. For example, items designed to gather demographic information often fail to acknowledge emic differences in populations. Asking for the age of a participant may appear to be a straightforward and simple question that anyone could answer, but this is not always the case. In some cultures date of birth is not considered a relevant piece of information, and the approximate age of a child is more often calculated by references to agricultural times of the year or other events. Educational attainment, income, employment and marital status are all demographic questions for which special awareness of within-culture perspectives are needed (Braun & Mohler, 2003).

Before addressing the particulars of best practices in translating an instrument, it is necessary to determine whether concept equivalence exists. It is possible that concepts that are emic – that is, unique to the source culture – render the line of research invalid in another culture. A study conducted in New Zealand using the SF-36, a widely used instrument that measures health related quality of life, found that emic views of health among a Maori sample did not discriminate between physical and mental health even though the items were written to differentiate between the two concepts (Scott et al., 1999).

Normative Comparability. Rules about disclosure of information between in-groups vary greatly. What can be said about particular subjects can be greatly influenced by what is perceived to be appropriate within an in-group. Religious preferences, political views and information related to personal matters are usually topics for which willingness to respond varies by culture. When interviewers and respondents share the same ethnicity, the willingness of the respondents to disclose, and hence the validity and the reliability of the data provided, increases (Marin & Marin, 1991).

Semantic Comparability. Semantic equivalence is related to the degree to which terms in the translation connote the same meaning in translation as in the source language. Achieving semantic comparability in questionnaire translation has been considered by some to be the most difficult step in the translation and adaptation of questionnaires. And, perhaps the greatest challenge to functionally equivalent questionnaires is the lack of equivalent markers or terms for words or concepts that do not exist in the target language or culture.

Research describing troublesome areas in test or questionnaire translations often focuses on this type of comparability issue (Arias et al., 1999; Lange, 2002). In the U.S., for example, the foster system trains and pays parents to care for children on a temporary basis. Collecting information about foster children is quite challenging for cultural, linguistic, and semantic reasons, however, as the term ‘foster child’ has no comparable translation in many other countries.

Trying to keep language simple has been long been advocated in writing survey questions (Peterson, 2000) but due to the absence of equivalent terms simplicity does not always equate to brevity. Terms that need to be translated often have either a more specific or a broader meaning in the target language. As a result, additional information needs to be presented. For example, the word ‘youth’ in English is often translated as ‘niño (child)’ in Spanish. However, the word ‘niño’ denotes someone twelve years old or younger, and is therefore not the optimal choice. To arrive at the closest meaning to the word it is necessary to use the equivalent of ‘young person’ and specify a range of ages included within the term. Doing this requires either adding more items to the instrument (Smith, 2003) or more words to an item.

In cross-cultural research, translating a word to its precise meaning is further complicated by regional variations in the particular language. Marin and Marin (1991) suggest that to avoid offending any given subgroup within a culture the standard or most neutral version of the language should be used. If regional variations are needed to accommodate a subgroup, both the standard term and the specific variation should be presented.

2.2 The impact of demographic variation across cultures

Educational level is often overlooked when designing cross-cultural questionnaires. U.S. literacy data reveal that a large number of adults can only read at elementary levels (Weech-Maldonado et al., 1999), with recent immigrants from Guatemala and El Salvador reporting less than a secondary education.

Flores et al. (2002) report that translations are written 'at an inappropriately high reading level for the target populations' and suggest that research instruments be validated by Spanish speaking families with 'poor and low-literacy populations'. In addition, when addressing different generations within one-culture, accommodations should to be made to meet their particular needs. During the 2000 Census, Chinese translations used traditional Chinese. Pan (2003) reported that during that census younger generations of immigrants were not able to read the traditional Chinese characters, reducing the pool of participants in this event.

Although the relationship between literacy levels and the ability to follow skip patterns in cross-cultural research has not been studied, familiarity with questionnaires is likely to influence the ability to follow instructions and respond to items. Forms literacy is an important element to consider when constructing a questionnaire. Navigating through an instrument presents a challenge to many foreign born respondents. Lack of exposure to ordinal scales and to multiple choice answers may require training for the participant (Lange, 2002).

2.3 Approaches to translation and cultural adaptation of instruments

The last twenty years have seen great changes in the methodology of instrument adaptation. There are many approaches to translation and cultural adaptation of research instruments. Popular criteria for evaluating translation techniques include the level of comparability achieved when translating an instrument and the cost involved in the task. Besides these criteria, research has focused on the number of individuals involved in completing a particular translation task or on the steps that should be taken to ensure the translation work is done accurately. The most frequently used techniques are:

- *Simple direct translation*, the translation of a document by a single individual from its original version to another language, is the easiest and cheapest method used. The most obvious drawback of this approach is that it does not offer ways to verify that the translation is true to the original and that different aspects of the language used are the most appropriate.
- *Translation with back translation* is a technique that requires at least two individuals. In this method, a bilingual individual translates the instrument from the original or source language to the target language. A second individual, without knowledge of the contents of the original document, translates the document back to the source language. The original and back translations are compared and differences between the documents are addressed. The translation/back translation approach allows the researcher to better assess that the translation is true to the original. The cost of this method rises considerably due to multiple steps and time needed for the process to be finalized.
- *Translation by committee* requires a group of individuals to arrive at a consensus about the best translation of an instrument after a series of steps.

Harkness (2001) provides an original framework for understanding instrument translation procedures. Within this conceptual framework, existing instruments are either adopted or adapted. Adopting involves directly translating the existing instrument into the target language with little attention to culture. Adapting requires reviewing an existing questionnaire for cross-cultural appropriateness in the target language, adjusting the source language as necessary, then translating the questionnaire into the target language.

Translators can employ sequential, parallel or simultaneous development of a new survey instrument. Sequential development refers to instrument development procedures in which the instrument is designed and pretested in the source language only. Once the instrument is finalized or 'locked', it is translated. Thus, the sequential development process does not take into account cultural or linguistic issues until after the finalized source language instrument is translated. Sequential development is comparable to adopting an existing instrument. The approach is efficient, but minimizes attention to cross-cultural issues. The parallel development process incorporates input from all target cultures during instrument development. An instrument is designed and pretested in the source language with a multicultural team. As with sequential development, the instrument is locked prior to translation. The most culturally adaptive method of developing an instrument is the simultaneous approach. Survey designers using simultaneous instrument development seek to create more than one version of an instrument using decentring, a process in which cultural appropriateness is given equal importance in all languages. The

instrument is designed and pretested simultaneously in multiple languages (Harkness, 2001). This process offers tremendous benefits for developing valid questions that are appropriate within multiple cultures. However, it tends to be expensive and time consuming, and it may produce ambiguous questions. The need arises for a lower cost methodology to simultaneously develop new questionnaires and to adapt, rather than adopt, existing instruments.

3 Question Appraisal Systems

As the use of cognitive pretesting and evaluation methods became a matter of course for questionnaire development in the 1980s and 1990s, the need arose for low-cost methods to assess the cognitive characteristics of questions that might lead to response error. Methods such as cognitive interviewing, behaviour coding, interview observation, and embedded question wording experiments were invaluable, but were at times found to be expensive and time-consuming. Question appraisal methods were designed to meet the demand for a more cost-effective way to systematically assess cognitive problems with instruments.

A questionnaire coding system developed by Lessler and Forsyth (1996) was designed based on Tourangeau's model of the question response process. The four cognitive processes in Tourangeau's model are comprehension, memory retrieval, judgment, and response selection (1984). Lessler and Forsyth found that, compared to other pretesting methods, their coding system was less expensive to implement and identified similar problems. Forsyth and Hubbard (1992) had previously had similar findings when they validated a questionnaire appraisal system by using cognitive think-aloud interviews. In 1999, Willis and Lessler developed the QAS-99, a tool based on previous questionnaire appraisal systems. The QAS-99 was an effort to 'assist questionnaire designers in evaluating survey questions, and in finding and fixing problems, before questions "go into the field",' (1999). The QAS-99 differed from Lessler and Forsyth's system in that it had significantly fewer codes as a means to decrease the difficulty of the coding activity and improve inter-coder reliability. The QAS-99 was designed not to replace other interactive cognitive methods, but to provide questionnaire reviewers and developers a tool for systematically reviewing and improving survey instruments.

The QAS-99 is comprised of seven steps by which a user evaluates each item in a questionnaire. Each step maps to a specific stage in Tourangeau's question-response model (1984). The seven steps and their associated question-response stages are displayed in Table 1.

Table 1 Steps in the Question Appraisal System (QAS-99)

QAS Coding Step	Stage in Tourangeau's Question-Response Process
Reading: Determine if it is difficult for the interviewers to read the question uniformly to all respondents.	Occurs prior to question-response process
Instructions: Look for problems with any introductions, instructions, or explanations from the respondent's point of view.	Comprehension
Clarity: Identify problems related to communicating the intent or meaning of the question to the respondent.	Comprehension
Assumptions: Determine if there are problems with assumptions made or the underlying logic.	Comprehension
Knowledge/Memory: Check whether respondents are likely to not know or have trouble remembering information.	Memory/Retrieval
Sensitivity/Bias: Assess questions for sensitive nature or wording, and for bias.	Judgment
Response Categories: Assess the adequacy of the range of responses to be recorded.	Response Selection

Source: Tourangeau, 1984

An additional eighth step provides an 'other' category for problems that cannot be assigned to one of the previous seven steps (Willis & Lessler, 1999).

4 Expanding the QAS-99

As the volume of multilingual surveys has increased over the past several years, survey designers have recognised the advantages of working more closely with translators throughout the questionnaire development process. A best practice used by the authors of this research involves collaborating with translation experts as early as possible in the instrument development process. Recognizing that establishing cross-cultural and multilingual validity is vital to reducing survey response error, a natural next step was to expand the QAS-99. The system was revised to include codes to allow for a systematic evaluation of a questionnaire for problems that might emerge in translation and cross-cultural application.

Prior to expanding the QAS-99, common best practices synthesised recommendations from previous research (Brislin, 1993; Maxwell, 1996) to develop a series of practical guidelines to establish item and concept equivalence across languages. These guidelines are provided in Table 2 and Table 3. These practical guidelines form the basis of the new steps added to the QAS-99.

Table 2 Practical Guidelines for Developing Cross-Cultural Surveys

Use reference periods that are culturally relevant
Avoid making assumptions about knowledge that may only be applicable in the source culture (e.g. religion, sports, holidays, other customs)
Be aware that name formats vary
Avoid using pictorial information that may not be fitting to the population studied
Use seasonal and holiday references that are part of the targeted culture

Table 3 Practical Guidelines for Developing Questionnaires for Translation

Avoid using double negatives
Using short sentences of less than sixteen words
Employ active rather than passive voice
Repeat nouns instead of replacing them with pronouns
Avoid metaphor and colloquialism
Avoid adverbs and prepositions telling where or when
Avoid possessive forms if possible
Avoid sentences using the same verbs if the verbs are used to mean different actions
Establish measurement and weight equivalences for cultures using metric systems
Use nouns common to the culture
Be specific when using the word “you”
Remember gender specific references might create a need for longer sentence structures

Two new QAS steps incorporate criteria for evaluating questions for multilingual and cross-cultural administration:

- *Cross-cultural:* Assess questions for problems in the response process that may emerge when the instrument is applied to varied cultures.
- *Translation:* Identify areas that can be clarified for a more accurate translated instrument.

In addition, a step for cross-question problems was included (displayed in the Appendix). Step 8, Cross-Cultural Considerations, consists of seven codes. These are displayed in Table 4. Most of these seven codes address concept equivalence. Reference Periods (8a), Knowledge (8b), Measuring Units (8c), Assumptions (8d) and Response categories (8e) highlight concepts that tend to vary across cultures. For example, Assumptions (8d), can be used to identify emic population differences that may affect questions about religious

practices (such as in predominantly Christian or Muslim countries) or sports (such as European football versus U.S. football). Additionally, Name Format (8f) and Politeness (8g) address normative equivalence, such as the appropriate way to identify individuals and the perceptions of polite conversation within a culture.

Table 4 Cross-Cultural Codes

ID	Proposed Code	Definition
8a	Reference Periods	The use of seasons as a reference period might be ambiguous or uncommon. Consider converting into months. Date formats vary. Consider using words for the month to avoid misunderstandings (e.g., 15 April 2004 vs April 15, 2004).
8b	Knowledge	Knowledge may not exist. Respondent is unlikely to know the answer to a factual question because he/she not familiar with the source culture. A culture-specific example is health insurance in the US for respondents who originate from countries with nationalized health insurance.
8c	Measuring Units	Consider reporting measuring units in both the English system and the metric system.
8d	Assumptions	Consider revising culturally inappropriate assumptions, including statements related to: sports, drugs, foods, drinks, activities, meal time, music, family ties, holidays, religion, books, magazines, school system, health system, and history.
8e	Response Categories	There is no equivalent concept or rating scale in foreign language. Avoid rating scales with more than 5 categories.
8f	Name Format	Response categories lack a space for other types of names. For example, Spanish speakers use paternal last name as well as maternal last name. Consider other naming conventions.
8g	Politeness	Courtesy and politeness can differ between cultures. Consider adding a 'Please' before commands like, 'Do not include …,' 'Mark every …', 'List all …' Consider using 'could' instead of 'should' if possible. Some commands or instructions might be perceived as rude, and respondents could change their attitude towards participating.

In contrast, Step 9, Potential Translation Problems, is focused more on semantic equivalence. Table 5 displays the codes featured in Step 9. Step 9 features codes that help ensure that the words used in translation have the same or similar meaning as in the source language. Therefore, Idioms (9b) are identified as problematic because they tend to lack equivalent meaning in translation. Likewise, vague quantifiers such as Time Adverbs (9e) should be specified with time periods as much as possible to clarify meaning.

Table 5 Translation Codes

ID	Proposed Code	Definition
9a	Double Negatives	This type of construction is hard to translate and can easily cause misunderstandings in other languages.
9b	Idioms	Many idioms do not have an equivalent across languages.
9c	Acronyms	The acronyms have no meaning in other languages. Consider providing an explanation with the acronym.
9f	Term 'You' is not Defined	Need to define the word 'You' (i.e., plural, singular, feminine, masculine, formal, informal).
9h	Time Adverbs	Need to avoid adverbs in the use of time: recently, lately, usually. Time references might be understood somewhat differently between languages. Consider specifying time frame.
9j	No Equivalent Term or Concept	Consider including an additional explanation.
9m	Adjectives Modifying Other Adjectives	Using adjectives to modify other adjectives, (e.g. 'house warming party', which must be literally translated from English to Spanish as 'A party in celebration of the purchase of a home in which guests take presents for the new home owner') is an uncommon grammatical usage in languages other than English. Consider paraphrasing and clearly define each term.
9n	References Applicable only to English	Toll free numbers, Web sites, contact information, books and other references are only available in the source language. Consider verifying which services or references are available in the target language. Also consider using numbers instead of letters on phone numbers.

These two new QAS categories provide survey designers with a tool for facilitating multicultural instrument development within the source language prior to translation. The QAS-04 uses conceptual, normative, and semantic understanding of the target language to broaden an instrument's applicability. Moreover, the relatively low costs associated with applying the QAS meet the need for an affordable way to conduct parallel development of a survey for translation.

5 Directions for Future Research

The QAS-04 provides a valuable tool for incorporating parallel question development into the source language questionnaire development and pretesting process. At minimum, it provides monolingual questionnaire developers with a tool by which they can anticipate question problems that may only show up in cross-cultural administration or translation. It provides a mechanism for bringing translation staff into the questionnaire design process at an earlier stage. The QAS-04 should be evaluated in a test application with an existing

questionnaire. It may be useful to compare the results of a test application to actual questionnaire data and interviewer experiences. Such a comparison would provide information on whether the QAS identifies the same problems interviewers and respondents experience in the field. The QAS-04 should also be assessed for internal consistency and reliability of the coding scheme, in order to assess the feasibility of using the codes.

In addition, the QAS-04 should be validated with reliability testing between monolingual and bilingual coders. At this time it is unclear whether monolingual coders and bilingual coders are equally capable of implementing the QAS-04. Moreover, the QAS-04 should be comparatively applied to questionnaires in the target languages as well as in English. Comparing the results could be a mechanism for validating translations as well as assessing the effectiveness of the tool itself.

Ultimately, the tool will be validated by its usefulness in developing cross-cultural and multilingual questionnaires. Only repeated use can determine whether it makes the survey development process easier or more efficient. It is our hope that the QAS-04 will provide the survey research community with a technique for systematically incorporating parallel questionnaire development into surveys that must be fielded in multiple languages and/or multiple cultures.

References

Arias, I., E. Rodríguez, J. Padilla, N. González, and M. A. Rodríguez. 1999. "Translation to Spanish, Reproducibility, and Cross Cultural Adaptation of the Miller-Rahe Recent Life-Change Questionnaire." *Arthritis Care and Research* 12:153-170.

Berry, J. W. 1969. "On Cross-Cultural Comparability." *International Journal of Psychology* 4:119-128.

Braun, M., and P. P. Mohler. 2003. "Background Variables." Pp. 101-116 in *Cross-Cultural Survey Methods*, edited by Harkness, J. A. et al. Hoboken: John Wiley and Sons, Inc.

Brislin, R. 1993. *Understanding Culture's Influence on Behavior.* 2nd ed. Fort Worth: Harcourt Brace College Publishers.

Cronin, M. 2003. *Translation and Globalization.* London, New York: Routledge.

Douglas, S. P., and C. S. Craig. 1983. *International Marketing Research.* Englewood Cliffs: Prentice Hall.

Flores, G., E. Fuentes-Afflick, O. Barbot, O. Carter-Pokras, L. Claudio, L. Marielena, J. A. McLaurin, L. Pachter, F. R. Gomez, F. Mendoza, R. B. Valdez, A. M. Villarruel, R. E. Zambrana, R. Greenberg, and M. Weitzman. 2002. "The Health of Latino Children: Urgent priorities, unanswered questions, and a Research agenda." *Journal of the American Medical Association* 288:82-90.

Fowler, F. J. 1995. *Improving Survey Questions: Design and Evaluation.* Thousand Oaks: Sage.

Forsyth, B., and M. Hubbard. 1992. "A Method for Identifying Cognitive Properties of Survey Items." Pp. 470-475 in *Proceedings of the Section on Survey Research Methods, American Statistical Association*. Alexandria: American Statistical Association.

Harkness, J. A., F. J. R. van de Vijver, and T. P. Johnson. 2003. "Questionnaire Design in Comparative Research." Pp. 19-34 in *Cross-Cultural Survey Methods,* edited by Harkness et al. Hoboken: John Wiley and Sons, Inc.

Kumata, H., and W. Schramm. 1956. "A Pilot Study of Cross-Cultural Meaning." *Public Opinion Quarterly* 20:229-238.

Lange, J. W. 2002. "Methodological Concerns for Non-Hispanic Investigators Conducting Research with Hispanic Americans." *Research in Nursing and Health* 25:411-419.

Lessler, J. T., and B. H. Forsyth. 1996. "A Coding System for Appraising Questionnaires." Pp. 259-292 in *Answering Questions: Methodology for Determining Cognitive and Communicative Processes in Survey Research,* edited by Schwarz, N., and S. Sudman. San Fransisco: Jossey-Bass Publishers.

Li, R. M., P. McCardle, R. L. Clark, K. Kinsella, and D. Berch. 2001. *Diverse Voices: The Inclusion of Language Minority Populations in National Studies: Challenges and Opportunities.* Report published jointly by the National Institute of Aging, the National Institute of Child Health and Human Development, and the National Center on Minority Health Disparities. http://www.nichd.nih.gov/publications/pubs/diverse_voices.htm.

Locke, D. C. 1998. *Increasing Multi-Cultural Understanding: A Comprehensive Model.* 2nd ed. Thousand Oaks: Sage Publications, Inc.

Marin, G., and B. V. Marin. 1991. *Research with Hispanic Populations.* Newbury Park: Sage Publications, Inc.

Maxwell, B. 1996. "Translation and Cultural Adaptation of the Survey Instruments." Pp. 8-10 in *Third International Mathematics and Science Study (TIMSS) Technical Report, Volume I: Design and Development,* edited by Martin, M. O., and D. L. Kelly. Chestnut Hill: Boston College.

Pan, Y. 2003. Paper presented at the Annual Meeting of American Association for Applied linguistics. *Sociolinguistic Research and Survey Development.* Arlington, VA.

Peterson, R. A. 2000. *Constructing Effective Questionnaires.* Thousand Oaks, CA: Sage.

Scott, K. M., D. Sarfate, M. I. Tobias, and S. J. Haslett. 1999. *A Challenge to the Cross-Cultural Validity of the Sf-36 Health Survey: Maori, Pacific, and New Zealand European Ethnic Groups.* Seventh Conference on Health Survey Research Methods. Hyattsville, MD: U.S. Department of Health and Human Services Publication No. (PHS) 01-1013.

Smith, T. W. 2003. "Developing Comparable Questions in Cross-National Surveys." Pp. 69-92 in *Cross-Cultural Survey Methods,* edited by Harkness, J. A. et al. Hoboken: John Wiley and Sons, Inc.

Tourangeau, R. 1984. "Cognitive Science and Survey Methods." Pp. 73-100 in *Cognitive Aspects of Survey Methodology: Building a Bridge Between Disciplines,* edited by Jabine, T. et al. Washington: National Academy Press.

Weech-Maldonado, R., B. O. Weidmer, L. S. Morales, and R. D. Hays. 1999. *Cross Cultural Adaptation of Survey Instruments: The CAHPS Experience.* Seventh Conference on Health Survey Research Methods. Hyattsville, MD: U.S. Department of Health and Human Services Publication No. (PHS) 01-1013.

Willis, G., and J. T. Lessler. 1999. *Question Appraisal System: QAS-99.* Rockville: Research Triangle Institute.

Appendix: Questionnaire Appraisal System (QAS-2004)

- **STEP 1 – READING: Determine if it is difficult for the interviewers to read the question uniformly to all respondents or if the reading level is appropriate.**
  - **1a. WHAT TO READ:** Interviewer may have difficulty determining what parts of the question should be read.
  - **1b. MISSING INFORMATION:** Information the interviewer needs to administer the question is not contained in the question.
  - **1c. HOW TO READ:** Question is not fully scripted and therefore difficult to read.
- **STEP 2 – INSTRUCTIONS: Look for problems with any introductions, instructions, or explanations from the respondent's point of view.**
  - **2a. CONFLICTING OR INACCURATE INSTRUCTIONS,** introductions, or explanations.
  - **2b. COMPLICATED INSTRUCTIONS,** introductions, or explanations.
  - **2c. MISSING OR INCONSISTENT INSTRUCTIONS** for DON'T KNOW and REFUSED answers.
- **STEP 3 – CLARITY: Identify problems related to communicating the intent or meaning of the question to the respondent.**
  - **3a. WORDING:** Question is lengthy, awkward, ungrammatical, or contains complicated syntax.
  - **3b. TECHNICAL TERM(S)** are undefined, unclear, or complex.
  - **3c. VAGUE:** There are multiple ways to interpret the question or to decide what is to be included or excluded.
  - **3d. REFERENCE PERIODS** are missing, not well specified, or in conflict.
  - **3e. PASSIVE VOICE:** Question is written in passive voice. Active voice is clearer both in source language and in translation.
- **STEP 4 – ASSUMPTIONS: Determine if there are problems with assumptions made or the underlying logic.**
  - **4a. INAPPROPRIATE ASSUMPTIONS** are made about the respondent or about his/her living situation.
  - **4b. ASSUMES CONSTANT BEHAVIOR** or experience for situations that vary.
  - **4c. DOUBLE-BARRELED:** Contains more than one implicit question.

Appendix (continued)

- **STEP 5 – KNOWLEDGE/MEMORY: Check whether respondents are likely to not know or have trouble remembering information.**
  - **5a. KNOWLEDGE** may not exist: Respondent is unlikely to know the answer to a factual question.
  - **5b. ATTITUDE** may not exist: Respondent is unlikely to have formed the attitude being asked about.
  - **5c. RECALL** failure: Respondent may not remember the information asked for.
  - **5d. COMPUTATION** problem: The question requires a difficult mental calculation.
- **STEP 6 – SENSITIVITY/BIAS: Assess questions for sensitive nature or wording, and for bias.**
  - **6a. SENSITIVE CONTENT (general):** The question asks about a topic that is embarrassing, very private, or that involves illegal behavior. If question will be applied across cultures, it may be sensitive in some cultures but not others.
  - **6b. SENSITIVE WORDING (specific):** Given that the general topic is sensitive, the wording should be improved to minimize sensitivity.
  - **6c. SOCIALLY ACCEPTABLE** response is implied by the question. If question will be applied across cultures, social acceptability could vary.
- **STEP 7 – RESPONSE CATEGORIES: Assess the adequacy of the range of responses to be recorded.**
  - **7a. OPEN-ENDED QUESTION** that is inappropriate or difficult.
  - **7b. MISMATCH** between question and response categories.
  - **7c. TECHNICAL TERM(S)** are undefined, unclear, or complex.
  - **7d. VAGUE** response categories are subject to multiple interpretations.
  - **7e. OVERLAPPING** response categories.
  - **7f. MISSING** eligible responses in response categories.
  - **7g. ILLOGICAL ORDER** of response categories.
- **STEP 8 – CROSS-CULTURAL CONSIDERATIONS: Assess questions for inappropriate or ineffective cross-cultural references.**
  - **8a. REFERENCE PERIODS:** The reference period uses seasons, American MM/DD/YYYY format, or may be otherwise ambiguous or unusual in other cultures.
  - **8b. KNOWLEDGE** may not exist: Respondent is unlikely to know the answer to a factual question because he/she not familiar with the American culture. Example: health insurance.
  - **8c. MEASURING UNITS:** Measuring units are from English system. If surveying Latin Americans or Western European populations, the metric system should be used.
  - **8d. ASSUMPTIONS:** The question includes culturally inappropriate assumptions or graphics. All statements related to sports, drugs, foods, drinks, activities, meal time, music, family ties, holidays, religion, books, magazines, school system, health system, and history should be evaluated.
  - **8e. RESPONSE CATEGORIES:** There is no equivalent concept or rating scale in foreign language. Avoid rating scales with more than 5 categories.
  - **8f. NAME FORMAT:** Response categories lack a space for other types of names. Spanish speakers use maternal last name as well as paternal last name, and other cultures list the family name as the first name.
  - **8g. POLITENESS:** Courtesy and politeness can differ in other cultures. Consider adding a 'Please' before commands like, 'Do not include …,' 'Mark every …', 'List all …' Consider using 'could' instead of 'should' if possible. Some commands or instructions might be perceived as rude, and respondents could change their attitude towards participating.

Appendix (concluded)

► **STEP 9 – POTENTIAL TRANSLATION PROBLEMS: Identify problematic question characteristics.**

9a. DOUBLE NEGATIVES: This type of construction is hard to translate and can easily cause misunderstandings in other languages.

9b. IDIOMS: Many idioms do not have an equivalent in other languages.

9c. ACRONYMS: The acronyms have no meaning in other languages. Consider providing an explanation with the acronym.

9d. UNCLEAR USE OF THE TERM 'YOU': 'You' not defined as plural, singular, feminine, masculine, formal, informal – a necessary step for translation.

9e. TIME ADVERBS: Question or response categories use adverbs to describe time, such as recently, lately, usually. Consider specifying time frame with number of days, weeks, etc.

9f. NO EQUIVALENT TERM OR CONCEPT in foreign language. Text may require an additional explanation.

9g REFERENCES APPLICABLE ONLY TO ENGLISH: Toll free numbers, Web sites, contact information, books and other references are only available in the source language. Consider verifying which services or references are available in the target language. Also consider using numbers instead of letter on phone numbers.

9h. ADJECTIVES MODIFYING OTHER ADJECTIVES: Using adjectives to modify other adjectives, (e.g. 'house warming party', which must be literally translated as 'A party in celebration of the purchase of a home in which guests take presents for the new home owner') is an uncommon grammatical in usage languages other than English. Consider paraphrasing and clearly define each term.

► **STEP 10 – CROSS-QUESTION: Look for cross-question problems in the entire questionnaire.**

10a. QUESTION PLACEMENT. The questions are not positioned in the most adequate section or order.

10b. DATA COLLECTION MODE: Sensitive question may be more effective if it was administered through another data collection mode.

10c. INCONSISTENCY WITH OTHER QUESTIONS: Wording, or response categories lack consistency.

10d. CONTENT OF PREVIOUS QUESTION AFFECTS MEANING: Does the content of the previous question or section affect the interpretation of the current question.

10e. SKIP PATTERN PROBLEM: Skip pattern is illogical or inadequate.

10f. FORMATTING: Layout or formatting is difficult to follow.

► **STEP 11 – OTHER PROBLEMS**

11a. QUESTION CONTAIN IRRELEVANT INFORMATION

11b. INAPPROPRIATE READING LEVEL

11c. OTHER PROBLEMS

USING A MIXED INTERNATIONAL COMPARABLE METHODOLOGICAL APPROACH IN A EUROPEAN COMMISSION PROJECT ON GENDER AND ENGINEERING

*FELIZITAS SAGEBIEL**

1 Introduction

The paper will focus on the European Project WomEng[1] „Creating Cultures of Success for Women Engineers“ (*www.womeng.net*). The project started on November 2002 and will last until end of October 2005. Participant members are universities and non profit women's engineering associations from seven countries (UK, France, Germany, Austria, Finland, Greece and Slovakia). In four so called work packages (wp) WomEng combines a strong quantitative with a complex qualitative methodology. There are two project parts: education and profession of engineering. In the first part (now completed) questions of choice of degree courses (work package 2) are connected with questions about experiences, satisfaction and dissatisfaction of students (work package 3) and questions of organisational cultures of degree courses (work package 4). A special work package focuses on methodology[2]; others on dissemination and coordination.

The article will demonstrate international and multi methodological comparisons on the issues of gender in engineering education as well as difficulties and possibilities of management of such a complex investigation. The first focus will be on work package 4 (wp4) because of restricted space and because it has been under the German responsibility.

---

* I want to thank Dipl. Soc. Wiss. Jennifer Dahmen for revsion and giving helpful feedback to the article. For translation help I thank student Jenia Bouxman, for formal adaption student Shirin Reinhard.

1 Besides Felizitas Sagebiel (University of Wuppertal, Germany) Christine Waechter (IFF/IFZ Graz, Austria), Maureen Cooper (University of Stirling, UK), André Beraud and Jean Soubrier (INSA, Lyon, France), Anne-Sophie Genin (ENSAM, Paris, France), Päivi Siltanen (Witec, Finland), Dora Kokla (EDEM, Athens, Greece), Oto Hudec (Technical University Kosice, Slovakia) are working together in WomEng. The coordination lies by Yvonne Pourrat (CDEFI, Paris, France).

2 Coordinated by Anne-Sophie Genin (France) and Oto Hudec (Slovakia).

2 Research Questions and Methodology

The question of choice of methodology in empirical research has been always depending on research questions and hypotheses. Research questions in WomEng had been structured in 3 work packages mentioned. The hypotheses for analysis of women students, adapted to organisational culture of engineering degree courses, were taken from state of art of European and non European research. The operational definitions of research issues were done in connection with state of the art using different quantitative and qualitative methodological measures.

2.1 Hypotheses

Out of state of the art it was clear that a masculine organisational culture can not be explained by a single aspect but by a complex setting of different cultural characteristics of engineering education. Recruitment measures and welcome events seem to be the first step to offer a changed image of engineering degree courses.

Second interdisciplinary curricula combining engineering with some non engineering subjects together with people oriented teaching methods (e.g. team and group work) should attract more female students. By this measure female students should be more attracted and feel more comfortable while studying.

From North America the single sex learning environment was one alternative cultural setting to prove. As in the European partner countries possibilities for single sex studying in higher education do not exist at the moment, this hypothesis could only be proved in Germany where some single sex model projects exist in a few universities of applied sciences.

The minority situation of female students in engineering was one focus to look at and ask if it was evaluated negatively or positively, and if females would possibly meet hostile attitudes from their male colleagues and/or from teachers. From Australian research one hypothesis was taken, looking at faculty in departments of engineering education as possible and necessary change agents for innovations.

Another hypothesis was that the enduring masculine image of engineering in society is reflected in departments of engineering degree courses and this could be a barrier to female students. The possible conflict between identity as women and identity as engineering student, taken out of literature, should be proved about its effects.

The study atmosphere is constructed by environment and social relations. And even more, masculine jokes and stories characterise the study culture. These characteristics may lead to feelings of isolation because of minority status and female students could experience marginalisation. On the other hand a controversial hypothesis was that young female students would feel comfortable in engineering degree courses and integrated in organisational culture.

2.2 Definition of work package and selection of methods (Berting, 1979)

Basis for development of methods was the overall methodology, prepared in proposal and discussed and decided in project meetings. In the beginning of the project the leader of work packages and the different teams preferred specific quantitative or qualitative methods on the basis of proposal and the tasks formulated there. Starting at first with the quantitative questionnaires for engineering students out of wp2 perspective this instrument was enlarged to work package 3 and 4 issues to avoid several similar investigations with different samples of questions. But, by this way the questionnaires got larger and after all took nearly one hour to fill.

On the other hand qualitative instruments were preferred from leaders from wp3 as well as from wp4, work packages with tasks which had been formulated to get more complex information and a deeper understanding of interdependences. While discussing and changing of methods these instruments were used also for understanding the underlying reasons for choice or non-choice of engineering degree courses (wp2) and attitudes for drop out or persistence (wp3).

To get known the institutional possibilities and barriers for an innovative women friendly engineering culture, the proposal for wp4 planned besides quantitative questionnaires for students three types of qualitative methods: expert interviews, participant observation and document analysis (homepage analysis).

So for example for the task to gather information about the culture of engineering departments most of the qualitative methods have been constructed and data have been collected. To get different perspectives engineering students and faculty of departments and universities were asked. Faculty interviews with representatives for degree courses, expert interviews with members of steering committee and officials from equal opportunity office have been done. The perspective/view of students came in from individual interviews with persistent and non-persistent students and especially from focus groups with female and male students. Focus group discussions have been experienced being a helpful method in all partner countries, which made focus groups with students (Austria, France, Germany, UK).

In summary, the description, analysis and interpretation of culture of engineering departments is based on non-reactive methods like homepage-analysis and participant observation as well as on more subjective methods of expert interviews and focus groups, which reflect students and faculties point of views and attitudes. As a non reactive method 35 homepage analyses of the investigated institutions and degree courses from 6 partner countries – Slovakia did not have to do this method – have been done. Participant obser-

vation has been done in 4 partner countries (Austria, France, Germany, UK). The filled criteria catalogues had been analysed by one student from the German team, who wrote his diploma thesis about this issue.

For the task to determine the impact of innovative engineering courses as part of wp4 the same instruments like those selected for description of organisational culture were used.

3 Quantitative Methodologies – The Survey

The quantitative methods have been mostly restricted to the first part of the project. Most of the time has been concentrated on construction of written questionnaires. For the design of the guidelines validity, reliability and sampling methods had to be taken into account.

100 female and male engineering students (with questionnaire 1 – Q1) in each country were compared with a group of 100 non-engineering students (science, social science and humanities, economics) (with questionnaire 2 – Q2). In each country characteristic institutions of higher education were chosen for investigation.

3.1 Construction of questionnaires (Porst, 1998)

In WomEng several drafts of questionnaires have been made. At first the wp2 leaders prepared a draft of questionnaires. This method was intended to serve for getting data about choice of degree courses and steps to decision. Full of questions about heritage and connection of ancestors to engineering professions, it focused on reasons for decisions or non- for an engineering degree course. This draft was discussed partly very controversial on the first project meeting. After this meeting it took some time with misunderstandings on what was agreed and what issues should be included in the questionnaires. Because of coordination problems some partners who had not participated so much in this discussion, started investigation without the final version. On the basis of pre-tests, hypotheses and ideas to be included the questionnaires were changed several times. The language, in which discussion took place, was mostly English. Only between Austria and Germany German communication was possible and was practiced. Except the meetings discussions were made by emails. After all, contents and questions of all work packages were integrated in the questionnaires.

The final version was worked out in English language by a native speaker and a partner from Scotland, UK. These final drafts of Q1 and Q2 had to be translated again in national languages to be given to the students for filling. The translations were organized and done by national teams, but usually not controlled by language specialists.

After all, the questionnaires had to be prepared in a form to allow easy statistical analysis. For this task the Slovakian statistician controlled the questionnaires. Several possibilities in accordance to time and money were discussed even the possibility to take a private company as subcontractor for statistical preparation of data, but this idea was given up.

3.2 Sampling methods and practice per country

Two *standardised questionnaires* for engineering students in comparison to non-engineering students were prepared and were offered to 50 female and 50 male students as a control group in all 7 European partner countries. Sampling of degree courses was oriented on the basis of lower, middle and higher percentage of female students depending on national statistics, choosing the most from degree courses with the lowest number.

For reliability of the samples in different countries three comparable groups in different percentages should be taken from degree courses, referring to the number of women in them. For example, mechanical engineering or computer science should be taken for a very low percentage of women, surely depending on the national situation of partners. As example for high percentage of women bio-something or civil engineering (without architecture) were possible. Third to the samples of worst and best situation according to the percentage of female students there should be taken also an example with an average percentage of women. Because of national variations not in all countries the same degree courses were chosen. This sampling allowed internal control of engineering degree courses, if there were differences in choice, satisfaction and studying atmosphere. On the other hand it gave a basis for generalisation of results.

In the non-engineering control group 100 students (50 female and 50 male) should be chosen, 40% students in natural sciences, 20 % in social and human sciences and 40 % in economics.

In each country characteristic institutions of higher education should be chosen for investigation. The number and characteristics of chosen institutions should include their history, tradition, localisation and culture of higher engineering education. In Germany, for instance, aspects of federal system in higher education were included.

In Germany for the selection of good practice in engineering degree courses the opportunity to investigate a single sex model of industrial engineering degree course was taken. This example offered the possibility to learn how an organisational culture could be changed by an innovative degree course. Because in the other partner countries no such changes initiated by any innovative engineering degree courses have been described, there is no systematic international comparison possible.

3.3 Validity and reliability measures

In literature about adequate methodology for international cross-cultural and interdisciplinary research several equivalence problems are discussed (Simmet-Blomberg, 1998: 292-344), which had to be solved. First of all the term "engineer" had to be clarified. It was agreed upon that all students/persons with an engineer degree should be taken for comparison, even though we knew that different systems of secondary and higher education in different countries as France, UK and Germany lead to different professional competencies and positions.

For getting valid answers for issues of wp3 (satisfaction) and wp4 (organisational culture) students should have been advanced enough in their studies to be able to appreciate these questions. Similar sampling criteria had been formulated for non-engineering students.

Several equivalence problems had to be taken in account:

(1) Content equivalence, which means functional, conception and categorical equivalence: e.g. women's situation in EU partner countries which has been one background for the project, was appreciated to be equivalent in western European countries, but in comparison to Slovakia different. As a result of 45 years of communist political, social and cultural system there was expected a quite different gendered labour division in the professional sphere. So in Slovakia more women in engineering and science were expected in comparison to western European countries, while at the same time in the private sphere traditional labour division was still expected.

(2) Different cultural sensibilities to research measurement should be coped by communication between EU partners.

(3) Language and translation problems had to be solved.

(4) In respect to sampling methods there exists a conflict between casual and controlled sampling. While the first serves reliability the second allows intercultural comparison. In WomEng controlled sampling methods had been chosen.

(5) Equivalence of definition: Even if the chosen degrees and degree courses in European partner countries were not the same, there were different criteria for control (see above).

For *reliability* of results from quantitative questionnaires control groups are most important. In WomEng male engineering students as well as female and male non-engineering students should be taken in account according to the key moments for female students to decide to become engineers or not, go on with studies or drop out. Sampling of non-engineering control group had to take in account different national structural and institu-

tional possibilities for choosing a degree course. So for example in UK and France students have to decide at a very early stage of life depending on their exams in maths and physics and this pre-decides, what is open for them. In France the success in preparatory classes decides who can become an engineer – only the best ones can. On the contrary in Germany students are free to study anything after finishing with mature.

Even though *pre-tests* (Porst, 1979) should be made in national languages to control general understanding of the questions, not all national teams made these pre-tests. In countries which worked with pre-tests of the questionnaires problems of understanding were sent to coordination of questionnaires for revision.

3.4 Implementation of survey

The quantitative and qualitative sampling has been carried out based on the second overall methodology of the project. The first calculations and thoughts for the sample were made on the base of 200 distributed questionnaires per target group. With this more diversification would have been possible, which afterwards had to be skipped because of less total numbers.

An overall criterion for choosing or not choosing a university was that the considered institutions must offer special activities to recruit and inform girls. The speciality of Germany, which offers single-sex degree courses in engineering degree courses, was from high interest too. The University of Applied Sciences in Stralsund with the women's degree course Industrial Engineering was chosen as example for good practice.

The implementation strategy varied from country to country even though there were the same guidelines for all. Some countries mailed the questionnaires to institutions to be spread by some persons while others travelled to the selected institutions and spread the questionnaires by project team personal who often watched the filling of questionnaires being able to answer questions if there were any. Whereas in France and Germany most of the Q1 questionnaires were handed personally or with close connections to faculty teaching in engineering degree courses, in other countries like Greece and Austria the questionnaires were mailed partly with problems of getting back the responses in necessary numbers. The non-academic Greek and Finish teams had more problems to get contact to reference persons in universities. For the French team, working in engineering schools themselves it was harder to reach non-engineering students, for Germany it was likewise easy to spread Q1 and Q2 and get them back. Because the guidelines have been implemented differently, one cannot decide what difficulties in getting filled questionnaires were due to country specialities and what due to different handling.

Interesting is that in Germany, where many problems had been expected before starting because of the very liberal higher education system, the team did not meet any serious problems getting enough questionnaires back and sticking to the guidelines. The most prominent reason was probably the very thoroughly preparation in management from reference persons in every degree course. For students, who took part in the survey, small presents like pencil and chocolate were spread. The implementation of investigation was concentrated on 2-3 months.

Problems were mentioned, that questionnaires were very long, some items were misunderstood, not adapted to the question. Others were not adapted to the national context, e.g. women special recruitment programs do not exist in France. Another problem was the French idea to force engineering students to return Q1, while this practice was a way, which never would have been possible in the German system. So, as the French team sees volunteering as a source of bias, in Germany you only could rely on volunteer students.

For analysis questionnaires results were entered manual in most cases with the help of a specific soft ware.

A problem was that while prolonged time for construction of questionnaires semesters for students to be asked had finished in several partner countries, and so the time for starting the survey was later than first planned.

4 Qualitative Methodologies – Completion of Survey Data

The quantitative questionnaires were combined with a number of different qualitative methods from which the researchers hoped to get a deeper insight in interdependencies. Each of the following methods should not be seen detached from the other ones and has its own possibilities. To prove validity and reliability of the results, they can be compared among each other and also work in a supplementary way, e.g. the results of the quantitative interviews can be checked by focus group discussions.

For the qualitative part of the investigation the methodology of the wp3 and 4 were the base. Most of the qualitative methodologies were used for wp4 and constructed by the German team in cooperation with other partner countries. The results have been the base for reports for dissemination, which sum up organisational cultures and innovative structures of degree courses with a low percentage of women (Sagebiel & Dahmen, 2005).

4.1 Guidelines and sampling for qualitative methods

Students were qualitatively interviewed as individuals and in focus groups to see how much their individual experiences correspondent with each other. Focus groups should allow to prove if group dynamics can further the perception of and talking about gender

discriminating aspects of study life, to exchange attitudes of the image of engineering, and how to overcome eventually masculinities of education in engineering. Website analysis of women friendly and women aversive measures in advertising engineering degree courses was done to describe and compare the culture at first glance. Male and female faculty were interviewed with semi-structured expert interviews about the practice of their institution/department in advertising, teaching, advising, mentoring, working atmosphere, eventual marginalization or friendships, image of engineering, attitudes towards single sex education in engineering etc. Members of steering committees and equal opportunity officers were asked about political decisions in engineering education, for example the drop-out-situation and measures of change. Non-participant observations of departments and teaching situations allowed a different perspective on the engineering culture.

All qualitative methods should be done in the same institutions, universities and departments as chosen for Q1.

Interview guidelines

For all interviews a similar method was used as a guideline, combining open with partly closed questions, completed with estimations along with Likert scale.

Students interviews

For complementary of quantitative questionnaires there were about 10 *guided interviews with female students* about their choice of degree course, experiences and satisfaction with study life, content, teaching methods and atmosphere and knowledge about reasons for drop out (most of all work package 3 issues). For work package 4 relevant issues were experiences and attitudes they met during studying. For comparison with questionnaires results a similar sampling was chosen, taking 3 students from degree course with a very low percentage of women and 2 from a degree course with a high percentage of women. The five female students, who had dropped out, should be taken from a degree course with a very low percentage.

Faculty interviews

Faculty as representatives for degree courses but not responsible persons – this choice was made to avoid social desirable answers – could give information out of teacher's perspective and they could be asked about their attitudes and estimation of female students, the department culture and possible barriers (prejudices). So faculties were asked about their ideas to make engineering degree courses more attractive as well as what they still practiced on self-advertising. Teaching methods, system and organisation of advice and men-

toring and practical experiences with gender differences were of interest, especially what teaching methods they favour and what they think female students would prefer. Description of working atmosphere with Likert scale from competitive, supportive, hierarchic, communicative to traditional was asked and if they believe it will be more supportive for men or for women. The same was asked for the students' working atmosphere. Social network and students' integration in faculty's meetings was a further issue. Special problems of female students in a degree course with a low number of women were asked and if they heard about those problems in mentoring and advising hours. Faculty in degree courses with a high percentage of females should give hints for a possible good practice and how to change more traditional structures. Attitudes about single sex education were asked to test their acceptance of changing possibilities. Treatment of females, possible mobbing and knowledge about dropping out and special reasons of female students were interview parts, especially for measuring dissatisfactions (work package 3). How they estimate the image of engineering in society and if this image is reflected in the culture of department was an issue for work package 4. The semi-structured questionnaire for expert interviews was similar to the faculty.

All interviews should be taped and transcribed afterwards. They should be analysed and interpreted first on the national level and put in summaries and sent to work package leader to be included in cross-cultural comparative reports.

Guideline for focus discussion groups

Looking at state of the art, it seemed not to be an easy job to get information on gender sensitive issues in the engineering field, where the main culture is characterized by perfectionism, seeing weaknesses and problems as not social desirable. In a group situation it seemed more probable to get beneath the superficial level. When discussing the methodological design on international conferences and presenting the feared problems to get valid information on gender issues in engineering degree courses, women scientists proposed focus or discussion group as a method to get hidden opinions and attitudes, which are more difficult to get in individual interview.

The aim of *focus discussion groups* is generally to get closer to the understandings and views of participants on certain issues. Special attention had to be paid to the use of focus groups together with surveys: the focus groups could be used for testing results of the surveys or could provide the issues which will be tested with the questionnaires. Focus discussion group is an important method because in the focus groups you talk to several people at the same time and participants talk to each other and can compare their experiences and attitudes. Maybe it came out that two people see the same thing in different

ways. These differences are very interesting because they could provide information which had not been thought of before.

After discussion of several possibilities to have different guidelines for female and male groups it has been decided on basis of time and money to have one guideline with special questions for women and men differently to allow comparison of attitudes of female and male students. Both groups should be close to the final exams to be more aware of gender differences in comparison to first semester students. Sampling should be made from a degree course with a low number of female students. The male focus group should be taken from an equal degree course to allow comparison.

There have been two parts, one biographical sheet, especially for wp2 with data about relatives and their influence besides some other demographic data (secondary education, sex, age, nationality) and objective study information. The second part contains questions about issues on study background, quality of social network in connection with study life, study atmosphere, how it feels, image of engineering, if it corresponds to department culture. Last but not least there were special questions for females and males. Female students were asked about how they were treated in comparison to male students, about their possible role models, their opinion about the women's recruitment programs and if they would prefer to have more women in their degree course. Male students were asked about opinions about female students in their degree course and how they appreciate girls' recruitment programs, if they think that females feel alone sometimes and if they as male students would prefer to have more females in the course and if they think women are treated equally to men.

Size of the groups should be about 5 participants. The discussion should last up to 2 hours and be videotaped.

Guideline for observation

For wp4 open *participant observation* (Warwick, 1973) was chosen because with this method it was possible to evaluate the study environment as well as people acting and reacting in everyday situation. At the same time this method could control eventually social desirability of interviews as reactive methods.

Data from *participant observations* of co-operative structures and teaching styles (frontal lessons, teamwork, projects) in studying and laboratory situations in traditional engineering faculties and those with innovative degree courses helped to determine the impact of the latter. Lectures of different subjects should be chosen for observation (subjects having a strong image and subjects having a soft image). Observed lectures should be given by

men and by women. Besides visual characteristics of department and gender/diversity special information on boards etc. could be observed by visiting engineering departments. Observation can't be seen as a method which is limited to 'seeing', of course it's also 'hearing', 'feeling', 'talking' and 'reading'. Even though participant observation (overt and covert) is not the most reliable research method, it offers the possibility to study a process in action and it is easier to take note of non-verbal behaviour of the people being observed.

Guideline for website analysis

The *website analysis* as a type of *document analysis* has been used for investigation of different dimensions: text to analyse the written text, graphics to analyse pictures and the colours used, usability to analyse whether all relevant texts are available, whether all sites can be shown easily (no broken links), whether there is the chance of communication to get further information etc. To combine all these dimensions a criteria catalogue has been constructed which had to be filled during surfing on the website. Website analysis should give information on integrated internships, welcome meetings, mentoring-, equal opportunity-, gender mainstreaming-, diversity-programs, life long learning possibilities and re-entry programs.

Even though the importance of this method seemed to be not so high in some countries looking at the number of students who looked at it for information as answering in the quantitative questionnaire, the future development will increase the use of homepages as source for information.

4.2 Implementation of guidelines

Sampling

The German sample consisted of five institutions of higher education, which were chosen to include different types of universities as well as partly regional specialities from Eastern and Western Germany. The sample constructed for questionnaire survey with engineering students (Q1) was taken too for the qualitative methods. The questionnaire for non-engineering students (Q2) was distributed at four the following universities. As one institution for good practice for organisational culture of engineering degree courses a model single sex degree course in industrial engineering installed in the University of Applied Sciences in Stralsund has been chosen. One comparable institution was the Technical University of Applied Sciences in Berlin, in which students as individuals and in focus groups, faculty and experts were asked.

Because of the different school systems and decision processes and entrance criteria for engineering degree courses for France and UK the sampling for questionnaires for control groups has been different in comparison to guidelines and other partner countries. Both had many problems getting Q2 in return. In Slovakia one institution has been chosen for all qualitative interviews and the team point to the aspect that this allowed intra-institutional comparison of students', faculty' and experts' answers. The selection was aimed to obtaining of comprehensive overview at the problem at the same university from selected groups (female successful students, non persistent female students, faculty staff member and steering committee member). A similar sampling was taken in Austria. Sampling has been implemented differently, taking one institution (like Austria and Slovakia) or different universities (like Germany and France).

Field work

To find five engineering students for two focus groups (male and female ones separately) out of degree courses with a small number of female students was not so easy, especially female focus groups were not easily to build, when there was only one women in each semester or deepening course. Austria used a snowball system to get enough students for focus group. This strategy did not work effectively in Germany, where the selected universities were situated in long distance to the home university and several research methods had to be finished in a few days.

Participant observation was clear and easy to follow the indications. While in France it was not so easy to find a teacher volunteer who allowed his or her lecture to be observed, in Germany for example there have not been any problems to get allowance. The timing for observation was important, especially to be not too late in the semester.

Website analysis made no big problems, but asking about the feeling of an atmosphere from a photo and working with alternatives like "warm" and "supportive" seemed for some teams too subjective. Another point was to recognize a gender discriminatory language. In German it is easy, when only the masculine form is used as a norm. In English one can only recognize discriminatory elements, if in gender unspecific cases the reflexive pronouns are used in masculine form only: he, his, him etc.

4.3 Analysis of qualitative data

For wp4 the input from partner countries was collected depending on the methods taken to get results in two steps. The first step was to gather raw material from all partners (we got less from Greece and Finland) on base of summaries along guidelines which had been prepared by work package 4 coordination and discussed and changed on work package leader meetings and per email.

For wp4 deliverables a common structure has been prepared by German team. This structure was discussed and changed by partners on a work package leader meeting. For helping to fill the structure a table with all empirical methods together with the number of questions and items to be integrated in the structure had been prepared and spread. The German team asked for two national summaries along with the necessary two reports for dissemination. At the same time a common structure was given, and for operational definition a table, which included all questions from different methods – quantitative and qualitative –, detailed structured for the final reports. This table constructed to help to fulfil the task nevertheless disturbed some partners who were not accustomed to work with tables in there professional life. It worked only after there had been given a practical example of national reports together with the detailed structure.

5 Problems of Cross-Cultural Comparison (Harkness, 1998; Simmet-Blomberg, 1998)

In summary our experience with quantitative and qualitative methods in a gender focused European research project on engineering shows three levels of problems of cross-cultural comparison:

1st Problem of comparison of different cultures

Having experience in cooperation out of the former European Commission Project INDECS (*www.INDECS.uni-wuppertal.de*) the RTD-research project WomEng was not so overloaded with intercultural misunderstandings.

Some questions were depending on country specificities and did not make sense in all countries. For example in French engineering schools women's welcome days or girls recruitment programs are not known. So these questions were inappropriate in France.

There were some questions which were not political correct in one country, but these questions differed from one country to another. So diversity questions are not political correct in France and UK whereas questions about income are not political correct in Germany for instance.

Looking back to the difficulties in international cooperation one prominent point was the different working styles depending on disciplines in combination with nationality. In research there is a necessity to overcome the separated worlds by communication and to find common definitions and solutions how to investigate the issues adequately in all participating countries. Implicitly too was how the politically engaged research was legitimized in comparison to so called scientific distance.

Disparate secondary and tertiary educational systems together with the problem to compare particular societies which are mostly non-comparable, are the backgrounds from which young females decide their studies. These decisions are made with different perspectives for study and professional life. Nevertheless if you speak of Europe and this should be meaningful you have to make compromises and try to construct comparability to overcome particular perspectives.

But the first problem was that the different disciplinary and cultural backgrounds of the scientific workers led to very different preparations of the empirical work. The sampling concept besides of the different numbers for degree courses with high and low percentage of women students was practiced very differently from one country to another. In Germany for example there was a trial to get data from different regions and represent a bit the federal system of education. Also different types of higher education institutions for engineering education were taken into account: technical universities, universities for applied sciences and comprehensive universities.

There were some translation problems[3] and partly the final guideline, which was used, was not exactly the same. For example some translations from English to German did not meet the same connotation and made comparison problematic after all.

2nd Methodological problems

Methodological problems are partly at the same time due to cultural differences. Another important point was the underestimation of time to be needed for statistical analysis of questionnaires which was finished short time before final deadline for project reports.

If cross-cultural comparison should be done in time one need data from all countries to a special date to put them together for systematic analysis and reports. At first the project partner agreed upon the guidelines for quantitative questionnaires. So, one could expect to get cross-cultural comparable data more or less if one neglect translation problems from common English questionnaires to the national languages of partners.

But one problem was not taken into account, which seemed crucial after all: the schedules seemed to be of different obligation depending on national characteristics, type of the partner organisation etc. It has been an important experience that a clear and definitely formulated letter to all partner teams to get all reliable and valid information for international comparison did not lead to success every time. So the definition of deadlines was not the same in partner countries, partly along with the different roles in research. Delays

3 Complexity of those translation problems are described in Harkness, van de Vijver & Johnson, 2003.

of country specific raw data and summaries and delay of statistical analyses culminated so that for example reports on organisational culture in large parts have not been comparative analyses, especially not on quantitative results of the questionnaires with students. And for this empirical part so much preparatory work had been done.

Some tables integrated in guidelines for interviews disappeared from one country to another, and these questions were not answered therefore and could not be compared afterwards.

The implementation of qualitative interview guidelines were differently too. While some teams took the guideline as definitely others took it only to ensure that none of the important topics were left out.

So after all not all data were produced in every country, partly because of the lack of money, but mostly out of other reasons. The outcome for the quantitative part of the project was besides others that too small numbers in cells for special items existed which did not allow special statistical measures and tests.

3rd Problems in connection with gender research

International construction of a quantitative questionnaire is full of adventures, especially in a European project on gender issues, working together with a mixed sex partner team out of different disciplines as well as traditions/non-traditions of feminist thinking. Out of perspective from a women's studies researcher it has been somewhat strange to implement a women's studies project in a team of mostly non-feminist researchers – women and men. This meant that they were not at all familiar with thirty years enduring discussions of gender issues and theory. The implicit question was if gender is a category as others or if it is a social construction in different European countries, defined by gendered division of labour.

In methodological language the different situation of women in partner countries could be seen as equivalence problems of different gendered labour division in European societies (see 3.3).

During discussion of questionnaires several times questions were skipped because they openly transferred gender prejudices and argued that this was not political correct, asking those questions in another country like Germany.

One very different aspect of department culture in France versus Germany and Austria in connection with gender was special recruitment measures for female students. Whereas in Germany and Austria it is a question of quality of departments and degree courses, and the

German and Austrian team took these measures as criteria for good practice institutions, in France students and faculty did not know anything about these, and when the interviewees understood what meant, they answered this differentiation by sex should not be and would not be political correct.

In respect to single sex education cross-cultural comparison is not possible because in partner countries those models do not exist.

One inherent problem getting true answers in interviews and focus groups has to do with social desirability. Especially about discriminating characteristics there is a tendency to get unbiased answers. For example "willingness of verbal account" with jokes and stories as team characteristic – often told about in literature – perhaps was not openly talked about, because it is not political correct in most European countries at the moment.

6 Conclusions

To further international comparable methodological approach it is necessary to communicate possible cultural differences and prejudices to overcome them. Methods should be developed in communication with all scientists working together in the project, and in this doing formal commitments and informal communication channels should be used, even by emails. All persons engaged should be flexible and sensible enough to recognize strength and weaknesses of partners to compensate, when it is necessary to do successful serious research. These elements come together with more methodological characteristics of comparable research, like cultural diversity and problems of terms, vocabulary, translation, inconsistencies, of measurement differences (Harkness, 2005 on an International Methodological Workshop in Kosiçe, Slovakia).

References

Beraud, A., and J. Soubrier. 2003. „Are Women the Future of Engineering? WOMENG – A European Research on Women and Engineering Education." Pp. 312-318 in *Sefi Conference – Global Engineer: Education and Training for Mobility.*

Berting, J. 1979. "A Framework for the Discussion of Theoretical and Methodological Problems in the Field of International Comparative Research in the Social Sciences." Pp. 137-156 in *Problems in International Comparative Research in the Social Sciences,* edited by Berting, J., and F. Greyer. Oxford, New York: Jurkovich, Ray.

Connell, R. W. 1999. *Der gemachte Mann: Konstruktion und Krise von Männlichkeiten.* Opladen: Leske und Budrich.

Etzkowitz, H. et al. 2000. *Athena Unbound. The Advancement of Women in Science and Technology.* Cambridge: Cambridge University Press.

Faulkner, W. 2000. "The Power and the Pleasure: How does Gender 'stick' to Engineers?" *Science, Technology, & Human Values* 5(1):87-119.

Harkness, J. A. (Eds.). 1998. *Cross-Cultural Survey Equivalence.* Mannheim: ZUMA-Nachrichten Spezial, No.3.

Harkness, J., van de Vijver, F. J. R., and Johnson, T. P. 2003. "Questionnaire Design in Comparative Research." In *Cross-Cultural Survey Methods. Wiley Series in Survey Methodology,* edited by Harkness, J. A., F. J. R. van de Vijver, and P. P. Mohler. New Jersey: John Wiley and Sons.

Hudec, O. et al. 2004. "Women as Engineering Students in Slovakia." *Slovak Sociological Review* 6(36):561-576.

McLean, C. et al. 1996. "Masculinity and the Culture of Engineering." Pp. 32-41 in *Third Australasian Women in Engineering Forum*, edited by University of Technology Sydney.

Porst, R. 1998. „Im Vorfeld der Befragung: Planung, Fragebogenentwicklung, Pretesting." *ZUMA-Arbeitsbericht* 98/02.

Sagebiel, F. 2004. *Masculinities' als Barrieren für angehende Ingenieurinnen in Europa?* Konferenzdokumentation 30. Kongress von Frauen in Naturwissenschaft und Technik 20.05.- 23.05.2004 in Winterthur. (printing).

Sagebiel, F., and J. Dahmen. 2005a. "Masculinities in Organisational Cultures in Engineering Education in Europe. Results of European project WomEng." *European Journal of Engineering Education 2005* (printing).

Sagebiel, F., and J. Dahmen. 2005b. „Männlichkeiten' in der europäischen Ingenieurkultur im EU-Projekt WomEng." *Soziale Technik* 1(15):19-21.

Sagebiel, F., and G. Hoeborn. 2004a. „Kultur in den Ingenieurwissenschaften – Standard und Abweichung" in *Standardabweichung. Dokumentation 29,* edited by Käthe und Clara e.V. - Verein zur Förderung von Frauen und Mädchen in Naturwissenschaft und Technik. Kongress von Frauen in Naturwissenschaft und Technik 29.05.-01.06.2003 in Berlin. Kirchlinteln.

Sagebiel, F., and Hoeborn, G. 2004b. „Männlichkeiten in den Ingenieurwissenschaften Europas – Theorie, Empirie und Veränderungspotenziale." *Journal* 16:27-37.

Saris, W. E., and M. Kaase (Hrsg.). 1997. *Eurobarometer: measurment Instruments for Opinions in Europe.* Mannheim: ZUMA-Nachrichten Spezial, Bd.2.

Simmet-Blomberg, H. 1998. *Interkulturelle Marktforschung im europäischen Transformationsprozess.* Stuttgart.

Stock, J. 1998. *Potentiale und Dimensionen der Wissensgesellschaft – Auswirkungen auf Bildungsprozesse und Bildungsstrukturen.* Delphi-Befragung 1996/1998, Basel.

Wächter, C. 2003. *Technik, Bildung und Geschlecht.* München, Wien: Profil Verlag.

Wajcman, J. 1991. *Feminism Confronts Technology.* Cambridge: Polity Press.

Warwick, D. P. 1973. „Survey Resarch and Participant Observation: A Benefit-Cost Analysis." Pp. 189-203 in *Comparative Research Methods,* edited by Warwick, D. P. New York: Osherson.

CULTURAL VARIABILITY IN THE EFFECTS OF QUESTION DESIGN FEATURES ON RESPONDENT COMPREHENSION

TIMOTHY P. JOHNSON, YOUNG IK CHO, ALLYSON HOLBROOK, DIANE O'ROURKE, RICHARD WARNECKE & NOEL CHÁVEZ*

To identify the characteristics of common health survey questions that may be associated with cross-cultural variability in question comprehension, health survey interviews with respondents representing four distinct cultural subgroups in the United States (non-Hispanic White, African American, Mexican American and Puerto Rican) were analyzed via behavior coding. Using survey responses as the unit of analysis (n=13,514), nested within survey respondents (n=345) and survey questions (n=42), hierarchical linear modeling (HLM) was employed to examine the effects of four questionnaire design features on cultural variations in question comprehension difficulties. Question response format, question length, question reading level and level of abstraction were each found to have main effects on respondent comprehension. Respondent culture was found to moderate the effects of response format, question length and reading levels. Several question design strategies that reduce overall comprehension difficulty also increase cross-cultural disparities in this regard.

1 Introduction

Respondent culture is now generally understood to influence the comprehension and interpretation of many health survey questionnaires (D'Andrade et al., 1972; Jenkins, 1988; Johnson et al., 1997; Meredith & Siu, 1995; Morse & Morse, 1988; Teresi et al., 2001; Warnecke et al., 1997). Culture-based variations in question comprehension may contribute to differential response artifacts that are erroneously interpreted as cultural differences or disparities in health beliefs, behaviors, and/or conditions (Johnson et al., 1996). Several

* This study was supported in part by Cooperative Agreement #U83/CCU508663 from the National Center for Health Statistics, U.S. Centers for Disease Control and Prevention.

approaches have been taken to address this problem in practice. A variety of new questionnaire translation strategies, for example, have been proposed in recent years to address cross-group differences in question comprehension (Harkness, 2003; Sperber, Devellis & Boehlecke, 1994). Several sets of question wording principles have also been proposed as guidelines for improving comprehension equivalence across cultural groups (Brislin, 1986; McKay et al., 1996; Smith, 2004). Little research, however, is currently available that can provide guidance to researchers regarding other survey question design features that might be useful in reducing comprehension variability when conducting health surveys in culturally heterogeneous environments, such as the United States.

Ironically, numerous question design features are known to be associated with respondent comprehension (Groves et al., 2004; Tourangeau, Rips & Rasinski, 2000). Several of these are believed to reflect question difficulty and/or complexity, including question length, reading level, abstraction level, and response format (Andrews, 1984; Bradburn & Miles, 1979; Bradburn & Sudman, 1979; Just & Carpenter, 1992; Knäuper et al., 1997; Laurent, 1972). The cross-cultural utility of these various question elements, however, have yet to be explored.

The purpose of this study is to investigate each of these common dimensions of health survey questions in order to assess any cultural variability in respondent comprehension that may be associated with each. To accomplish this, behavioral coding (Fowler, 1995) is applied to measure comprehension difficulties across 345 health survey interviews with a culturally diverse sample of U.S. respondents and a range of 42 survey questions. These survey data and behavioral codes offer to provide an assessment of the degree to which the effects of question design features do or do not operate in similar manners across race/ethnic groups.

2 Methods

With respondent consent, a total of 345 in-person laboratory interviews were tape-recorded. Using race/ethnicity as a proxy measure of respondent culture, four groups of respondents were examined: African Americans (n=86), Mexican Americans (n=101), Puerto Ricans (n=74), and non-Hispanic Whites (N=84) residing in the Chicago Metropolitan Area. Respondents were recruited via advertisements in local media and ranged in age from 18-53. The interviews averaged approximately an hour in length and all were conducted in English.

2.1 Survey questions

The survey instrument included 42 substantive health-related questions selected from national health surveys conducted in the United States. Among the surveys from which items were selected were the National Health Interview Survey (NHIS), the Behavioral Risk Factor and Surveillance Survey (BRFSS) and the National Household Survey of Drug Abuse (NHSDA). Items were selected to represent a variety of topics, question types and formats. The specific wording of all questions is available from the authors. Following each interview, respondents completed a brief inventory of demographic questions.

The 42 survey questions were classified along four dimensions: question length, reading difficulty level, response format, and abstraction level. The length of each question was measured by total number of words. The reading level of each question was graded using Flesch-Kincaid scores (Flesch, 1979). Three response formats were included: those asking for numeric values (e.g., number of times exercise, age first drank alcohol) (n = 13); those for which the respondent could answer "yes" or "no" (n = 9); and those employing vague quantifiers as response categories (e.g., "excellent-good-fair-poor," "strongly agree-agree-disagree-strongly disagree") (n = 20).

Three levels of abstraction were identified. Two of the authors independently classified each question as "most abstract," "somewhat abstract," or "least abstract." Results were subsequently compared and differences discussed and reconciled. Abstract items were defined á priori as those for which the major concept introduced by the question was not grounded in physical reality (n = 11 of the questions examined). Those items classified as "least abstract" were those for which the major concept introduced in the question was grounded in physical reality (n = 17 questions). The remaining 14 items were classified as "somewhat abstract."

2.2 Response coding

Audio-tapes were reviewed and respondent reactions to each of the 42 substantive survey questions were coded using a behavioral coding scheme previously reported (Oksenberg, Cannell & Kalton, 1991) and modified for this study. A graduate assistant who was trained and supervised by one of the authors coded a total of 13,514 respondent answers. A random sample of 24 tapes was coded by both persons, revealing an inter-rater agreement of 98.1 percent. Table 1 defines five specific respondent behavior codes that were classified as indicative of comprehension difficulty. An exploratory factor analysis was conducted to evaluate the dimensionality of these five behavior codes. All five behaviors loaded strongly on a single factor. Survey questions that elicited one or more of these behavior codes were consequently defined as producing comprehension difficulty for the respondent in question. Overall, comprehension difficulties were associated with 9.4 percent of the 13,514 survey responses analyzed.

Table 1 Respondent Behavior Codes Used to Represent Comprehension Difficulties

1. *Clarification (Unspecified):* Respondent indicates uncertainty about question, but it is unclear as to whether the problem is related to the construct or the context.
2. *Clarification (Construct):* Respondent asks for repeat or clarification of question, or makes a statement indicating uncertainty about question ***meaning*** (i.e., "what do you mean by depressed?").
3. *Clarification (Context):* Respondent indicates s/he understands the meaning of the construct, but indicates uncertainty about question meaning within the context of the question as stated (i.e., "what do you want to know about being depressed?").
4. *Clarification (time frame):* Respondent indicates uncertainty about the question's time frame.
5. *Clarification (rewording):* Respondent rephrases the question before answering.

Descriptive information for each question level and respondent level variable is presented in Table 2.

Table 2 Question and Respondent Variables

Variables	Mean	SD	Minimum	Maximum
Question Characteristics		Level 1 *(N=13,514)*		
Comprehension Difficulty	.09	.29	0	1
Abstraction Level				
Abstract	.27	.45	0	1
In-between	.35	.48	0	1
Concrete	.38	.49	0	1
Question Length	21.80	10.52	5	46
Readability Level	6.84	2.81	1.50	12.0
Response Type				
Numeric	.28	.45	0	1
Yes/No	.22	.41	0	1
Vague quantifiers	.50	.50	0	1
Respondent Characteristics		Level 2 *(N=345)*		
Education	3.58	1.16	1	6
Age	32.08	8.28	18	53
Gender (Male)	.51	.50	0	1
Race/Ethnicity				
White	.24	.43	0	1
African American	.25	.43	0	1
Mexican American	.29	.46	0	1
Puerto Rican American	.21	.41	0	1

2.3 Analysis

In order to estimate two-level hierarchical linear models, HLM6 (Raudenbush et al., 2004) was employed. A multi-level modeling strategy enabled an accounting of variance attributed to individual differences (i.e., responses are nested within subjects) and factors associated with individuals as well as questionnaire characteristics. As depicted below, a general conditional model is composed of two sets of equations: equation 1 at the response level, and equation 2 at the respondent level. Since the outcome variable (comprehension difficulty) is binary (problem=1; no problem=0), the model predicts the expected log-odds of the outcome at the first level using a logit link function (Hedeker & Gibbons, 1994). At the first level, expected log-odds of the comprehension difficulty [Log [P/(1-P)]] are predicted by questionnaire characteristics, including abstraction level, question length, reading level and response format. At level-2, intercept at level *1*(B_0) is modeled as a function of level-2 predictors (respondent demographics) controlling for the random variability (μ_0) across individual respondents.

Equation 1: Level-1 Model:

$$\text{Log [P/(1-P)]} = \beta_0 + \beta_1(\text{Most Abstract}) + \beta_2\,(\text{Somewhat Abstract}) + \beta_3\,(\text{Question Length}) + \beta_4\,(\text{Reading Level}) + \beta_5\,(\text{Yes-No Responses}) + \beta_6\,(\text{Vague quantifier Responses})$$

Equation 2: Level-2 Model:

$$\beta_0 = \gamma_{00} + \gamma_{01}\,(\text{Education}) + \gamma_{02}\,(\text{Age}) + \gamma_{03}\,(\text{Male}) + \gamma_{04}\,(\text{African American}) + \gamma_{05}\,(\text{Mexican American}) + \gamma_{06}\,(\text{Puerto Rican}) + \mu_0$$

In addition, the effects of level-1 factors are predicted by race/ethnicity to explore interaction effects between level-1 question characteristics and the respondent's race/ethnicity. These effects are estimated in a second model that employs equation 1 and equation 3 as follows:

Equation 3: Level-2 Model to examine interaction effects:

$$\beta_0 = \gamma_{00} + \gamma_{01}\,(\text{Education}) + \gamma_{02}\,(\text{Age}) + \gamma_{03}\,(\text{Male}) + \gamma_{04}\,(\text{African American}) + \gamma_{05}\,(\text{Mexican American}) + \gamma_{06}\,(\text{Puerto Rican}) + \mu_0$$

$$\beta_1 = \gamma_{10} + \gamma_{11}\,(\text{African American}) + \gamma_{12}\,(\text{Mexican American}) + \gamma_{13}\,(\text{Puerto Rican American})$$

$$\beta_2 = \gamma_{20} + \gamma_{21}\,(\text{African American}) + \gamma_{22}\,(\text{Mexican American}) + \gamma_{23}\,(\text{Puerto Rican American})$$

$$\beta_3 = \gamma_{30} + \gamma_{31}\,(\text{African American}) + \gamma_{32}\,(\text{Mexican American}) + \gamma_{33}\,(\text{Puerto Rican American})$$

$$\beta_4 = \gamma_{40} + \gamma_{41}\,(\text{African American}) + \gamma_{42}\,(\text{Mexican American}) + \gamma_{43}\,(\text{Puerto Rican American})$$

$$\beta_5 = \gamma_{50} + \gamma_{51}\,(\text{African American}) + \gamma_{52}\,(\text{Mexican American}) + \gamma_{53}\,(\text{Puerto Rican American})$$

3 Results

Table 3 presents HLM model results for the main effects of person level and question level characteristics on question comprehension, and Table 4 presents interaction effects between race/ethnicity and each question level indicators. Among the respondent characteristics examined, only race/ethnicity was found to be independently associated with question comprehension problems. Members of minority groups (i.e., African-American, Mexican-American and Puerto Rican respondents) were each more likely to express comprehension difficulties when compared with non-Hispanic white respondents. Respondent age, gender, and education were not related to difficulties.

Table 3 HLM Estimates of Main Effects of Individual and Question-Level Characteristics on Comprehension Difficulty

	Coefficient	(SE)
Effects of the individual characteristics		
Intercept	-2.98***	(0.26)
Education	-0.05	(0.04)
Age	0.01	(0.01)
Gender (Male)	0.05	(0.09)
Race/Ethnicity (Ref=White)		
African American	0.30*	(0.12)
Mexican American	0.50***	(0.12)
Puerto Rican American	0.40**	(0.14)
Effects of questionnaire characteristics		
Abstraction Level		
(Ref= Least Abstract)		
Most Abstract	0.73***	(0.09)
Somewhat Abstract	0.12	(0.08)
Question Length	0.01***	(0.00)
Reading Difficulty Level	0.07***	(0.01)
Response Format		
(Ref= Numeric)		
Yes/No	-1.28***	(0.10)
Vague Quantifier	-1.20***	(0.08)

*p<.05; **p<.01; *** p<.001.

Each of the four question characteristics was found to be independently associated with respondent comprehension. The effects of three were found to vary across race/ethnic groups. There was a main effect of level of question abstraction. Not surprisingly, those items classified as “most abstract” produced more comprehension difficulties, compared to items designated as “least” abstract. No differences were found between items classified as “moderately” and “least” abstract. The effects of question abstraction did not vary by race/ethnicity as shown in Table 4.

Table 4 HLM Estimates of Interaction Effects of Race/Ethnicity and Question Characteristics on Respondent Comprehension Difficulty

	Coefficient	(SE)
Effects of individual characteristics		
Intercept	-3.37***	(0.30)
Education	-0.05	(0.04)
Age	0.01	(0.01)
Gender (Male)	0.05	(0.09)
Race/Ethnicity (Ref=White)		
African American	0.65**	(0.24)
Mexican American	1.05***	(0.23)
Puerto Rican American	0.83**	(0.25)
Interaction effects between questionnaire characteristics and respondent race/ethnicity		
Abstraction Level (Ref=Somewhat Abstract)		
Most Abstract		
Intercept (White)	0.81***	(0.20)
African American	-0.48	(0.26)
Mexican American	0.08	(0.26)
Puerto Rican American	0.11	(0.27)
Moderately Abstract		
Intercept (White)	0.17	(0.16)
African American	-0.19	(0.23)
Mexican American	0.05	(0.22)
Puerto Rican American	-0.05	(0.22)
Question Length		
Intercept (White)	0.02***	(0.01)
African American	-0.00	(0.01)
Mexican American	-0.01	(0.01)
Puerto Rican American	-0.02*	(0.01)
Reading Difficulty Level		
Intercept (White)	0.13***	(0.02)
African American	-0.08**	(0.03)
Mexican American	-0.10***	(0.03)
Puerto Rican American	-0.04	(0.03)
Response Format (Ref= Numeric)		
Yes/No		
Intercept (White)	-1.67***	(0.20)
African American	0.80**	(0.27)
Mexican American	0.33	(0.27)
Puerto Rican American	0.30	(0.29)
Vague Quantifier		
Intercept (White)	-1.81***	(0.20)
African American	0.97**	(0.27)
Mexican American	0.60*	(0.25)
Puerto Rican American	0.72**	(0.26)

*p<.05; **p<.01; *** p<.001.

Question length was found to have a positive main effect on comprehension difficulty: as the number of words increased, so did the likelihood that respondents would express comprehension problems. The negative coefficient associated with Puerto Rican status (-.02) in Table 4 indicated that difficulties with question comprehension increased with question length among non-Hispanic whites at a greater rate than among Puerto Ricans. Recognizing that Puerto Rican respondents in general expressed more comprehension difficulties than whites, it would appear that white respondents are more sensitive to variations in question length than are Puerto Ricans.

The reading difficulty level of survey questions was also found to be positively associated with comprehension difficulty: increased reading level was associated with more question comprehension difficulty. Race/ethnicity again moderated the effects of this variable. The negative regression coefficients associated with African-American and Mexican-American identities indicated that members of these two cultural groups experienced less additional comprehension difficulty as the reading level of survey questions increased, compared to White respondents. Although not significant, the regression coefficient associated with Puerto Rican ethnic identity suggests a similar relationship pattern. That is, reading level seems to be more problematic for White respondents (see Table 4).

Both main and interaction effects were also found for question response format. Specifically, question response formats that asked respondents to provide a numeric response (e.g. number of physician visits) generated more overall comprehension difficulties when compared with response formats in which respondents were asked to answer using a set of predefined response options (including both "yes-no" and vague quantifier formats). In addition, race/ethnicity was found to moderate the effects that the vague quantifier response format had on question comprehension. In particular, compared to Whites, African-American, Mexican-American, and Puerto Rican respondents were each more likely to express comprehension difficulties when vague quantifier response formats were employed. In addition, when 'yes-no' response formats were employed, African-American respondents were more likely to express comprehension difficulties, compared with Whites. The model in Table 4 was re-estimated to examine race/ethnic group differences in comprehension difficulties when asked questions requiring a numeric response (results not shown). No differences were found across groups in the likelihood of expressing comprehension difficulties with the numeric response format.

4 Discussion

This research used behavioral coding protocols for survey interviews to confirm cross-cultural differences in respondent's ability to comprehend a set of health-related survey items. Non-Hispanic Whites, in particular, expressed comprehension problems with a smaller number of survey questions, relative to the three minority groups also interviewed. These differences remained after controlling for other demographic characteristics, particularly age and education, that one might also expect to find associated with question comprehension. As the survey questions included were all selected from national health surveys in the United States, it is most likely the case that they were developed by representatives of the country's dominant non-Hispanic White culture. As such, it is not surprising that respondents from this group in general had less difficulty interpreting these questions.

Main effects of the four question characteristics evaluated were also identified. Questions deemed to be most abstract appear to be more likely to elicit comprehension difficulties among survey respondents, compared to those classified as least abstract. Likewise, comprehension problems also increase with the reading level and length of survey items. Response formats are additionally linked to comprehension problems, with questions that request numeric estimates generating more comprehension problem-related behaviors, compared to items that provide sets of response options. These findings are consistent with much of the available literature regarding question-based sources of processing error in survey research (Groves et al., 2004; Tourangeau, Rips & Rasinski, 2000). Findings related to race/ethnic differences in comprehension difficulties that are linked to specific question design features, however, have not been previously reported. Comprehension differences across groups were detected in responses to three of the four question design features examined: response format, question length, and reading level.

What is most perplexing in reviewing findings related to question response formats is that those formats that generate the fewest overall comprehension difficulties (i.e., the 'yes-no' and vague quantifier formats) ironically also generate the largest cross-cultural disparities. The greatest variability across groups was found in response to questions employing sets of vague quantifiers. The ambiguities associated with this response format are well known (Tourangeau, Rips & Rasinski, 2000). It is thus perhaps less surprising that cultural differences in comprehension difficulties were found among those questions employing sets of vague quantifiers as response options. We believe the same culture-based processes associated with main effect differences in comprehension difficulties across groups may also be responsible for these differences. Just as survey questions and instruments are largely prepared by White middle-class researchers in the U.S., so too are the response categories

attached to many of these items. It thus seems reasonable to assume that the processes responsible for general race/ethnic comprehension disparities in these data are also highlighted among those questions employing sets of vague quantifier response options, which are almost by definition subject to a greater range of interpretations than simple 'yes-no' formats.

Interestingly, one of the few available recommendations in regards to constructing cross-culturally equivalent survey items is the advice to use dichotomous response options ('yes-no,' 'agree-disagree') whenever possible (Smith, 2004). Our findings, however, indicate that some cultural variations in comprehension difficulties are also found among health questions employing binary 'yes-no' response options. Interestingly, no cross-group differences were found in difficulty answering numeric response format questions. Although this format elicits the greatest overall level of comprehension problems, it appears to produce little cross-group variability. That is, all respondents find numeric response formats equally and most difficult to answer. Of course, for many epidemiologic applications, numeric question response formats are unavoidable. Investigators should nonetheless exercise caution when requesting numeric information from survey respondents.

We also found differential effects of question reading level on respondent comprehension across race/ethnic groups. Specifically, increasing reading level was associated with more comprehension difficulty among White respondents, compared to African Americans and Mexican Americans. When examined in conjunction with the main effects of race/ethnicity on question comprehension, these findings suggest that questions with higher levels of reading difficulty may actually serve to *decrease* the gap in question comprehension between White and minority respondents. Overall, Whites are less likely to express comprehension problems. When confronted with questions at higher reading levels, however, Whites are more likely to exhibit comprehension problem behaviors, bringing them closer to the levels experienced by minority respondents in response to questions at all levels of reading difficulty. It would again appear that those questions that eliminate cross-cultural variability in comprehension (i.e., those worded at a high reading level) are also those that are most difficult for all respondents to comprehend. A similar pattern was found in regards to the cross-group effects of question length. Whereas shorter questions produced fewer comprehension difficulties, race/ethnic differences were minimized for longer questions.

Overall, these findings suggest that well-founded efforts to improve the general comprehension of survey questions may in some cases have the unintended effect of increasing cross-cultural variability. Simplifying question length, reading level and response formats would appear to improve overall question comprehension at the cost of enhancing cross-cultural disparities. These findings are perhaps a legacy of the process by which knowledge of survey question design features has accumulated over the past 50 years in the U.S.

This research has largely over-represented White, non-Hispanic survey respondents (Payne, 1951; Schuman & Presser, 1981; Sudman & Bradburn, 1982) and it should perhaps thus not be surprising that findings are less generalizable to respondents from other cultural backgrounds. Clearly, more work is essential to identify question design features that effectively minimize comprehension problems in general and which also eliminate disparities in comprehension across cultural groups.

We wish to qualify our findings by acknowledging several important limitations. These include the fact that only a relatively small sample of health-related survey questions (n=45) were examined. Future work will need to evaluate larger and more diverse sets of survey questions. Second, behavioral coding has not been previously used for the purposes described in this paper. This methodology was initially developed as a method of evaluating interviewer performance (Cannell et al., 1975) and has been previously used both to evaluate the performance of survey questions (Oskenberg et al., 1991; van der Zouwen & Smit, 2004) and survey interviewers (Dykema, Lepkowski & Blixt, 1997). We believe, however, that this approach to assessing cultural variations in question understanding has good face validity and is an appropriate use of these methods. Our sample of respondents was relatively small and non-random, raising questions about the generality of our findings. Finally, we also note that findings should not be generalized beyond the U.S. context. Nonetheless, this work is based on more than 13,000 survey responses by members of four distinct race/ethnic groups. In addition, the findings are consistent with a growing body of research literature demonstrating the existence of cross-cultural variations in the comprehension and interpretation of survey questions. The development of methods and procedures for establishing the conceptual equivalence of survey measures across cultures should be encouraged as the U.S. continues to evolve into a more culturally heterogeneous society.

References

Andrews, F. M. 1984. "Construct validity and error components of survey measures: A structural modeling approach." *Public Opinion Quarterly* 48:409-442.

Bradburn, N. M., and C. Miles. 1979. "Vague quantifiers." *Public Opinion Quarterly* 43:9-101.

Bradburn, N. M., and S. Sudman. 1979. *Improving Interview Method and Questionnaire Design.* San Francisco: Jossey-Bass.

Brislin, R. W. 1986. "The wording and translation of research instruments." Pp. 137-164 in *Field Methods in Cross-Cultural Research,* edited by Lonner, W. J., and J. W. Berry. Beverly Hills: Sage.

Cannell, C. F., S. A. Lawson, and D. L. Hausser. 1975. *A Technique for Evaluating Interviewer Performance.* Ann Arbor: University of Michigan.

D'Andrade, R., N. Quinn, S. Nerlove, and A. K. Romney. 1972. "Categories of disease in American-English and Mexican-Spanish." Pp. 9-54 in *Multidimensional Scaling,* volume II, edited by Romney, A. K., R. Shepard, and S. B. Nerlove. New York: Seminar Press.

Dykema, J., J. M. Lepkowski, and S. Blixt. 1997. "The effect of interviewer and respondent behavior on data quality: Analysis of interaction coding in a validation study." Pp. 287-310 in *Survey Measurement and Process Quality,* edited by Lyberg, L., P. Biemer, M. Collins, E. de Leeuw, C. Dippo, N. Schwarz, et al. New York: John Wiley & Sons.

Flesch, R. 1979. *How to Write Plain English.* New York: Harper and Row.

Fowler, F. J. 1995. *Improving Survey Questions: Design and Evaluation.* Thousand Oaks: Sage.

Groves, R. M., F. J. Fowler, M. P. Couper, J. M. Lepkowski, E. Singer, and R. Tourangeau. 2004. *Survey Methodology.* New York: John Wiley & Sons.

Harkness, J. A. 2003. "Questionnaire translation." Pp. 35-56 in *Cross-Cultural Survey Methods,* edited by Harkness, J. A., F. J. R. van de Vijver, and P. Ph. Mohler. New York: John Wiley & Sons.

Hedeker, D., and R. Gibbons. 1994. "A random-effects ordinal regression model for multilevel analysis." *Biometrics* 50:993-944.

Jenkins, J. H. 1988. "Conceptions of schizophrenia as a problem of nerves: A cross-cultural comparison of Mexican-Americans and Anglo-Americans." *Social Science and Medicine* 26:1233-1243.

Johnson, T. P., D. O'Rourke, N. Chavez, S. Sudman, R. Warnecke, L. Lacey, et al. 1997. "Social cognition and responses to survey questions among culturally diverse populations." Pp. 87-113 in *Survey Measurement and Process Quality,* edited by Lyberg, L., P. Biemer, M. Collins, E. de Leeuw, C. Dippo, N. Schwarz, et al. New York: John Wiley & Sons.

Johnson, T. P., D. O'Rourke, N. Chavez, S. Sudman, R. B. Warnecke, L. Lacey, et al. 1996. "Cultural variations in the interpretation of health survey questions." Pp. 57-62 in *Health Survey Research Methods Conference Proceedings,* edited by Warnecke, R. B. DHHS Publication No. (PHS) 96-1013. Hyattsville, MD: National Center for Health Statistics.

Just, M. A., and Carpenter, P. A. 1992. "A capacity theory of comprehension: Individual differences in working memory." *Psychological Review* 99:122-149.

Knäuper, B., R. F. Belli, D. H. Hill, and A. R. Herzog. 1997. "Question difficulty and respondents' cognitive ability: The effect on data quality." *Journal of Official Statistics* 13:181-199.

Laurent, A. 1972. "Effects of question length on reporting behavior in the survey interview." *Journal of the American Statistical Association* 67:298-305.

McKay, R. B., M. J. Breslow, R. L. Sangster, S. M. Gabbard, R. W. Reynolds, J. M. Nakamoto, et al. 1996. "Translating survey questionnaires: Lessons learned." *New Directions for Evaluation* 70:93-104.

Meredith, L. S., and A. L. Siu. 1995. "Variation and quality of self-report health data: Asians and Pacific Islanders compared with other ethnic groups." *Medical Care* 33:1120-1131.

Morse, J. M., and R. M. Morse. 1988. "Cultural variation in the inference of pain." *Journal of Cross-Cultural Psychology* 19:232-242.

Oksenberg, L., C. Cannell, and G. Kalton. 1991. "New strategies for pretesting survey questions." *Journal of Official Statistics* 7:349-365.

Payne, S. L. 1951. *The Art of Asking Questions.* Princeton: Princeton University Press.

Raudenbush, S. W., A. S. Bryk, Y. Fai Cheong, and R. T. Congdon. 2004. *HLM 6: Hierarchical Linear and Nonlinear Modeling.* Lincolnwood, IL: Scientific Software International, Inc.

Schuman, H., and S. Presser. 1981. Questions and Answers in Attitude Surveys. San Diego: Academic Press.

Smith, T. W. 2004. "Developing and evaluating cross-national survey instruments." Pp. 431-452 in *Methods for Testing and Evaluating Survey Questionnaires,* edited by Presser, S., J. M. Rothgeb, M. P. Couper, J. T. Lessler, E. Martin, J. Martin et al. Hoboken: John Wiley.

Sperber, A. D., R. F. DeVellis, and B. Boehlecke. 1994. "Cross-cultural translation: Methodology and validation." *Journal of Cross-Cultural Psychology* 25:501-524.

Sudman, S., and N. Bradburn. 1982. *Asking Questions.* San Francisco: Jossey-Bass.

Teresi, J., D. Holmes, M. Ramierz, B. Gurland, and R. Lantigua. 2001. "Performance of cognitive tests among different racial/ethnic and education groups: Findings of differential item functioning and possible item bias." *Journal of Mental Health and Aging* 7:79-89.

Tourangeau, R., L. J. Rips, and K. Rasinski. 2000. *The Psychology of Survey Response.* Cambridge: Cambridge University Press.

van der Zouwen, J., and J. H. Smit. 2004. "Evaluating survey questions by analyzing patterns of behavior codes and question-answer sequences: A diagnostic approach." Pp. 109-130 in *Methods for Testing and Evaluating Survey Questionnaires,* edited by Presser, S., J. M. Rothgeb, M. P. Couper, J. T. Lessler, E. Martin, J. Martin, et al. Hoboken: John Wiley.

Warnecke, R. B., T. P. Johnson, N. Chavez, S. Sudman, D. P. O'Rourke, L. Lacey, et al. 1997. "Improving question wording in surveys of culturally diverse populations." *Annals of Epidemiology* 7:334-342.

INTERPRETING THE RESULTS OF CROSS-CULTURAL COGNITIVE INTERVIEWS

A Mixed-Method Approach

*KRISTEN MILLER, GORDON WILLIS, CONNIE EASON, LISA MOSES & BETH CANFIELD**

Cognitive interviewing is used to empirically assess the ways in which individuals mentally process and respond to survey questions that are presented in either written or auditory form, and is commonly applied at the questionnaire pretesting stage (DeMaio & Rothgeb, 1996; Snijkers, 2002; Willis, 2005). However, there is little consensus among practitioners regarding the standards or criteria that constitute high-quality cognitive evaluations (Snijkers, 2003). While some limited research within the evaluation literature pertains to interviewing technique, regarding issues such as specificity versus generality of probes (Foddy, 1998), or concurrent versus retrospective probing (Redline, Smiley, Lee, et al., 1998), few investigations have attended to the analysis of cognitive interview results (Willis, 2005). That is, how are findings from individual cognitive interviews to be used in order to make conclusions regarding the functioning of a survey question? This analytic deficiency has raised skepticism regarding the replicability, falsifiability and, ultimately, the validity of cognitive interview findings (Conrad, Blair & Tracy, 2000; Tucker, 1997; Willis, 2005).

Two strands of thought appear to shape the discussion regarding the quality of information produced by cognitive interviews. The first argues that if the method is to provide meaningful results, cognitive interviews must be standardized; only structured interviews can be systematically analyzed for unbiased results (Tucker, 1997). The second argues that it is the qualitative results of loosely structured interviews – those allowing for spontaneous or emergent probing – that lends strength to cognitive interviewing methodology (Gerber, 1999). Unlike regimented interviews, semi-structured interviews may capture contextual information that is essential to understanding the interpretive aspects of the question-response process.

* The authors would like to thank Jennifer Madans of NCHS for her support and very helpful comments.

Underlying this discussion is the age-old epistemological discourse that results in the pitting of quantitative and qualitative methodologies. It is not our position that only quantitative methods allow for objective and systematic analyses of cognitive interviews. To the contrary, qualitative methodology has produced a vast literature on the systematic analysis of textual data which aptly defends the validity of such an analysis (Denzin & Lincoln, 2000). We suspect, however, that cognitive interviewing methods – while advantaged by the semi-structured interview – can be further enhanced by quantitative analysis, that is, when a mixed-method approach is used (Tashakkori & Teddlie, 1998). A mixed method approach incorporates the contextual data derived from semi-structured cognitive interviews along with numerical coding of results. The combination of qualitative and quantitative analyses presumably augments the quality as well as the variety of information that can be obtained through cognitive interviewing.

Further, we suspect that mixed-method approaches will be especially useful within a vital emerging area: The application of cognitive interviewing techniques across culture and language. Although there are several challenges to quality in cross-cultural cognitive interviewing (Johnson, 1998), the most vexing may be a reliance on a purely qualitative and sometime impressionistic interviewing approach. Because cognitive interviewers typically are nested within cultural or language group (i.e., they can only conduct interviews in languages they can speak, and must therefore employ bilingual staff to conduct other-language interviews), it is not clear whether the results across subgroups represent differences between the cultures represented (or questionnaire translations), or whether they simply reflect stylistic differences between disparate cognitive interviewers. Especially for purposes of assessing the cross-cultural equivalence of survey questions, it is vital that subgroup variation be adequately interpreted, lest investigators be led astray by variance that is only imposed by the question evaluation procedure. Hence, we propose that cross-cultural investigations may benefit from a more structured approach than has typically been employed for purposes of question pretesting, in order to minimize error due only to the interviewer.

Warnecke, Johnson, Chavez, et al. (1997) report on the application of a system for coding cognitive interviewing outcomes, in a study involving racially and culturally diverse populations. However, they provided little information concerning the nature of their coding system, other than implying that it was useful in systematically disentangling the results of a large number of cognitive interviews, and their codes appear to relate mainly to question comprehension processes. For the current study we therefore developed an explicit coding system that emphasizes the full range of cognitive processes postulated to influence survey responding. Codes were based on a cognitively-oriented question-response model of the type described by Tourangeau (1984) depicting (1) the interview

subject's interpretations of key terms, (2) the subject's ability to retrieve information necessary to answer the questions; (3) decision processes used to modify or further process information, or to judge its adequacy; and (4) matching of the respondent's internal representation of their answer to given response categories. These codes were applied to the unstructured written interviewer notes, to produce the quantitative component of a mixed-method analysis of interviews. The paper will describe the findings related to the tested survey questions, using an analytic approach in which the coding and tabulation of results was supplemented by interviewers' open-ended text comments. Further, we examined a range of respondent characteristics other than ethnicity (e.g. gender, ethnicity, age, language) that might impact the question response process. Finally, the paper describes the strengths and weaknesses and methodological utility of this analytic approach.

1 Method

Sample. Sixty-seven cognitive interviews were conducted among (self-reported) Hispanic and Non-Hispanic (both White or Black) participants in urban (Washington D.C.) and rural/suburban locations (two locations in Northwest Ohio). Because of the socio-cultural focus, a relatively equal distribution of participants across socio-economic status, gender, age, ethnicity, as well as language and geographical location, was selected (see Table 1). Participants were recruited through newspaper ads, fliers, and by word-of-mouth. The D.C. area interviews were conducted in the Questionnaire Design Research Laboratory at the National Center for Health Statistics. Ohio interviews were conducted either in the participant's home or in a private room of a community facility. All participants were remunerated $35 after the interview.

Table 1 Northwest Ohio & DC Metropolitan Cognitive Interview Subjects

	Race/Ethnicity	Income	Education	Age	Gender
DC Metropolitan (English) 12 Participants	White = 7 Black = 5	11-20K = 4 21-30K = 1 31-50K = 3 51-80K = 1 61-80K = 0 81K+ = 3	Elementary = 0 Some High School = 2 H. S. Grad. = 4 Some College = 6	18-29 = 0 30-49 = 0 50-69 = 7 70+ = 5	Female = 7 Male = 5
NW Ohio (English) 20 Participants	White = 19 Hispanic=1 Black = 0	0-10K = 7 11-20K = 6 21-30K = 5 Unknown = 2	Elementary = 3 Some H.S. = 7 H. S. Grad. = 8 Some College = 2	18-29 = 1 30-49 = 6 50-69 = 8 70+ = 5	Female = 12 Male = 8
NW Ohio (Spanish, Spanish/English combination) 35 Participants	Mex. Am. = 17 Mexican = 14 Puerto Rican = 1 Hispanic = 1 Hisp. Am. = 1 Cuban = 1	0-10K = 9 11-20K = 9 21-30K = 6 31-50K = 7 51-80K = 4	Elementary = 5 Some H.S. = 12 H. S. Grad. = 13 Some College = 2 Unknown = 3	18-29 = 7 30-49 = 16 50-69 = 11 70+ = 1	Female = 20 Male = 15

Data Collection. The interviews were based on an interviewer-administered health survey questionnaire containing items selected from the NCHS National Health Interview Survey, or created anew, covering chronic conditions, cancer screening, diet, physical activity and demographic characteristics. All but one of the cognitive interviews of Hispanics were conducted in Spanish, and all Non-Hispanics interviews were in English. The instrument was translated from English by one of the authors. The cognitive interviews were semi-structured; along with the survey questions, the interview guide (protocol) consisted of several pre-scripted follow-up questions pertaining to participants' interpretations of key terms and overall comprehension of questions. These fixed probes ensured that this particular information was collected in every interview and could then be compared across all interviews. As a less standardized approach, interviewers were also instructed to inquire as to the ways in which participants constructed their answers to the survey questions, which further provides insight into potential sources of response error. These *emergent*, non-scripted probes helped interviewers make sense of gaps or contradictions in participants' explanations and provided contextual information needed to precisely define question problems. In turn, this open-ended information contributed to the development of a coding system for purposes of succinctly characterizing the results in quantifiable form.

Code Development. Two sets of numerical codes were developed from the cognitive interviews, a *problem set* and an *interpretive set*. The problem codes, based on the standard question-response model (Comprehension, Retrieval, Decision and Response), indicate situations in which tested subjects deviated from or were unable to fully negotiate stages of the question response process. Table 2 outlines this set of codes. It should be noted that problem codes do not necessarily reflect the presence of actual response error. For example, many participants were unfamiliar with the term *chronic obstructive pulmonary disease*, yet based on the total evidence obtained, appeared to respond correctly that they did not have that condition. Nevertheless, because these participants were unable to fully comprehend the question, a problem code of 1 was assigned in such a case.

Because problem codes were based on the response process model, most of the codes were developed prior to interviews. However, a few response problems were not anticipated in the initial schema (e.g. codes 5 and 7) and were added as interviews were being conducted. Consequently, refinement of codes occurred inductively, which ensured that all nuances of question problems were included within the coding schema. Though this process established a complete data set, as a new code was developed, all previous interviews needed to be re-checked for consistency – an extremely time-consuming endeavor.

Table 2 Question-Response Problem Codes

Comprehension	1	*Term*: Subject does not understand or know the meaning of specific words
	2	*Question*: Subject does not understand the question as a whole because of vagueness or complexity
Retrieval	3	Subject does not know (and never knew) the requested information
	4	Subject is unable to remember requested information
Decision	5	Subject is unable to make calculations necessary to arrive at the answer
	6	Question sensitivity or perceived negative reaction by subject
	7	Subject is unable to decide on a response
	8	Subject is found to estimate either too high or too low
Response	9	Response categories do not match subject's internal representation of the answer
----	0	No problems observed

In addition to response problems, codes were developed based on interpretive patterns. These codes were not necessarily directly error-related, but reflected variation in the ways participants conceptualized key terms, such as *health, mid-day meal, advice*, and *exercise*. Unlike the problem codes, these interpretive codes were generated entirely from an inductive process, that is, based on qualitative analysis of the interview text. After interviews were collected, patterns of interpretation were identified across participants, and each pattern was then assigned a numerical code. For example, for the general health question, *"Would you say your health in general is excellent, very good, good, fair or poor?"*, two themes regarding participants' interpretation of the word *health* emerged: 1) a predominately physical conceptualization of health and 2) a multi-dimensional conceptualization including physical, but also mental, emotional and/or spiritual health. The following interview passages illustrate the two interpretive themes:

Physical Health: Coded 1 (physical):

"I'd say very good because I don't have any diseases, but I could be in better shape… you know, I should exercise more and it would be good if I stopped smoking completely."

Multi-dimensional Health: Coded 2 (multi-dimensional):

"My health is very good because I feel happy most of the time. I have a few aches and pains, but overall I feel good and I have a strong connection to God."

2 Results

Several types of analyses could be conducted using the final coded data set. Most broadly, problems were tallied to indicate a general frequency of problems posed by each question. To illustrate the wide range in problem severity as indicated by the overall measure used, Table 3 depicts (for 10 of the 30 total tested questions) the percentage of participants experiencing at least one type of coded problem (questions not illustrated revealed values that were intermediate in severity).

Table 3 Percentage of Subjects Producing at Least one Type of Question-Response Problem Code (for a Sample of Tested Questions)

Tested question	% subjects with 1+ codes
1. Do you have emphysema or chronic obstructive pulmonary disease (COPD)?	92.3% (60/65)
2. Do you have congestive heart failure?	77.5% (31/40)
3. What is the total value of all financial assets that you own? Please include Individual Retirement Accounts (IRAs), 401k plans, stocks and bonds, mutual funds, certificates of deposit (CDs), savings accounts, or any other financial assets.	78.1% (50/64)
4. Do you have chronic bronchitis?	53.7% (36/67)
5. Did you have a midday meal yesterday?	32.1% (17/53)
6. Do you have diabetes?	22.7% (15/66)
7. When you use butter or oils for cooking or preparing your food, which of the following types do you use most often? 1) Butter, Margarine, Lard, or Shortening, 2) Olive oil or Canola oil, 3) Corn oil, Vegetable oil, Peanut oil, Soy oil, 4) Non-stick spray, 5) Don't use fat	22.4% (15/67)
8. Would you say your health in general is excellent, very good, good, fair or poor?	11.9% (8/67)
9. Did you eat any other meals or snacks yesterday? (Other than the meals you just told me about)	4.5% (3/66)
10. Which fruit, vegetables, salad or juice did you have for a snack yesterday?	0.0% (0/66)

As illustrated in Table 3, a question on COPD (Chronic Obstructive Pulmonary Disease) scored the highest of the examined questions in this assessment. Understandably, all of the presented problems for this question were definition-based; 92% of the participants could not provide a correct definition for this term. Again, from this analysis it is not clear how serious this problem may be for estimate accuracy, as many participants were able to respond in a way that appeared to be accurate, based on further probing (however, it should be noted that a small proportion, thinking COPD equates to heart problems or asthma, answered the question incorrectly). However, this analysis did indicate that the problem itself is uncomplicated and could be easily fixed with a clarifying phrase or accompanying definition.

A question on total financial assets, on the other hand, presented a more complex set of problems. Only 1.6% of the problems associated with the asset question were definition-based. The primary problem pertained to inadequate response categories (61,0%), specifically that there was no adequate response category for those who had few or no such assets. Other problems involved question complexity (13.5%), inaccurate estimation (6.8%), lack of knowledge (8.5%) and question sensitivity (5,0%). Unlike the COPD question, problems involved every stage of the response process, and consequently there appeared to be no simple or straightforward modification that can be implemented to improve the question.

Statistical analysis of problem codes. Arguably, the types of analysis presented immediately above may not provide anything that a traditional cognitive analysis would omit, as any credible cognitive evaluation presumably should reveal these types of problems. The most advantageous aspects of this type of coding analysis, however, are (a) the ability to produce a quantitative estimate of potential problem severity (in terms of percentage of interviews in which a problem was in evidence), and (b) to explicitly investigate potential for group variation, by determining if a problem is more likely to occur within a particular group, as opposed to evenly distributed across respondents. To determine if problems were systematically related to ethnicity or to other measured subject characteristics, cross-tabulations and logistic regression analyses were conducted, involving age, gender, ethnicity, income, and education. Logistic regression analysis was done in hierarchical manner, with ethnicity (Hispanic/Non-Hispanic) entered last into the model, so as to determine the unique contribution of this variable, controlling for the influence of other measured demographics. Because 12 items exhibited either ceiling or floor effects (defined as >90% interviews illustrating error, and <10%, respectively), 18 items were considered statistically analyzable.

The regression analysis determined that for these 18 items, ethnic group membership was the strongest overall predictor of problem code frequency ($p < .05$), with Hispanics generally experiencing more difficulties than Non-Hispanics (for 5 items), but with Hispanics seemingly having fewer problems for two other evaluated questions. Somewhat surprisingly, normally analyzed demographic characteristics, including gender, age, educational level, and income, had relatively weak effects: Gender produced no significant effects for any item; older respondents had somewhat more trouble than younger ones with a question on lifting and carrying; and subjects with lower income produced more codes for a question concerning coronary heart disease. Subjects with lower educational level produced significantly more codes for the question *"Do you have chronic bronchitis?"*: 75% of those participants who did not graduate from high school, as opposed to 30% of those with a high school degree, exhibited some type of problem with this question. An exami-

nation of the qualitative interviewing results reveals that the difficulty centered primarily around uncertainty of the word *chronic*; in particular, participants with lesser education confused episodes of *acute* and *chronic* bronchitis. As a follow-up analysis, 80% of the more educated participants (along with 100% of the less educated) were found to have experienced terminology problems in the chronic obstructive pulmonary disease question, suggesting that few individuals are familiar with terminology associated with chronic disease that is commonly used within health surveys.

Overall, effects as measured by summary problem codes were mainly related to ethnic group membership. Even these were not consistently unidirectional, however. Hispanics produced significantly more problems with a question on ever having cancer, and significantly less for one on combined household income. Five other questions exhibiting Hispanic/Non-Hispanic differences involved food and meal questions; data are depicted in Table 4.

Table 4 Percentage of Participants Having Response Problems, by Ethnicity

Tested question	Hispanics	Non-Hispanics
1) How many times did you eat red meat, including beef, pork, lamb, or lunchmeat, hot dogs or sausages made from beef, pork or lamb yesterday?	77.1% (27/35)	35.5% (11/31)
2) Did you eat a morning meal yesterday?	66.7% (24/36)	6.5% (2/31)
3) Did you eat a midday meal yesterday?	54.2% (13/24)	13.8% (4/29)
4) Did you eat an evening meal yesterday?	57.7% (15/26)	6.7% (2/30)
5) When you use butter or oils for cooking or preparing your food, which of the following types do you use most often? 1) Butter, Margarine, Lard or Shortening, 2) Olive oil or Canola oil, 3) Corn oil, Vegetable oil, Peanut oil, Soy oil, 4) non-stick spray, 5) Don't use fat	8.3% (3/36)	38.7% (12/31)

Again, qualitative analysis facilitates interpretation of these differences, and to pinpoint the character of the response problem. From the very beginning of Spanish language interviewing, it was clear that some translated survey questions caused interpretation difficulties for Hispanic subjects. That is, particular words were translated literally from English and, because of cultural differences, did not convey the same meaning. For example, the phrase *frijoles con chile* was intended to mean chili beans, but was interpreted by most Hispanic participants as beans with hot sauce. Additionally, some words varied by particular region (e.g., Puerto Rican Spanish uses *nami* for yam, while Mexican Spanish

uses *camote*) or were more formal forms of Spanish (e.g., the word *fiambre* for lunch-meat). Consequently, these terms were not always understood by Hispanic participants. It is this variety of translation problem that apparently accounted for the higher percentage of Hispanics experiencing problems with the red meat question.

Similarly, some words in Spanish consisted of more than one meaning and could easily be taken out of context. For example, the word *comida* can mean *meal, food*, and the name of a meal – like the English word for dinner. Consequently, the question *"Did you eat a morning meal?"* was translated as *"¿Ayer comío Ud. la comida de la mañana?"* but misunderstood by some Hispanic participants as *"Did you eat your dinner in the morning?"* This interpretive and translative issue accounts for a large portion of the ethnicity-based response problems regarding the meal questions, and is evidenced by the following exchange, which also illustrates how cognitive probing brings out an otherwise "silent misunderstanding," as termed by DeMaio & Rothgeb (1996):

Interviewer:	*Digame, Ayer comio usted la comida de la manana?*
	Tell me, did you eat a meal in the morning?
Participant:	No.
Interviewer:	*Y la manana para Usted, que quiere decir, que tanto tiempo, de que horas a que horas?*
	And morning for you, what does it mean, what time frame or from what hour to what hour is it?
Participant:	*Pues en la manana el desayuno es a las nueve.*
	Well in the morning el desayuno is at nine.
Interviewer:	*So, el desayuno, lo nombra el desayuno, es a las nueve?*
	So, el desayuno, you name it desayuno, is at nine?
Participant:	*Si, por que you no doy el que le dicen... .como le dicen... Braaq faat*
	Yes, because I don't serve, what they call… how do they say…Braaq faat
Interviewer:	Breakfast?
Participant:	*Si.*
	Yes.
Interviewer:	*No hace breakfast sino que hace desayuno?*
	You don't make breakfast, but you make desayuno?
Participant:	*Si, yo desayuno, asi estoy acustumbrada…doy mi desayuno y mi comida y en la cena como algo mas liviano.*
	Yes, I have desayuno, that is how I am accustomed… I serve desayuno, and then my comida and for cena, I eat something a lot lighter.

In this case the subject was initially asked about "comida de la manana," or as intended by the investigators, "food in the morning." However, she evidently interprets "comida" as the meal eaten after her morning meal (which is "desayuno"), and so indicates that no, she has not eaten her "comida" meal in the morning – and produces what, to the survey administrator, is an erroneous response.

A qualitative examination of the interviews also reveals why Non-Hispanic subjects, in comparison to Hispanics, were likely to experience problems with the oil questions. Many Hispanics reported using butter and lard to cook, and, consequently were able to provide an answer with little consideration. Non-Hispanic participants, on the other hand, were much more likely to use a variety of cooking oils and experienced trouble determining which type of oil they used most often. To provide an answer, they often needed to mentally recreate their cooking habits – which oil they used for which specific type of food – and then consider which oil was used most often.

Analysis of the interpretive codes, while not necessarily signaling potential response error, also illustrates how ethnicity impacts the question-response process, particularly question comprehension. For example, from examination of the problem codes, there appeared to be few problems in the general health question, *"Would you say your health in general is excellent, very good, good, fair or poor?"* Those problems that were identified were related to response categories – in which participants had difficulty matching their personal conceptualization of health within the provided categories. However, analysis of the interpretive codes reveal an intriguing finding: While most of the Non-Hispanic participants (77%) gave reports coded as conceptualizing health as a physical phenomena, most of the Hispanic participants (90%) used a comprehensive conceptualization of health, incorporating emotional and spiritual dimensions. It is not surprising that Hispanic participants – especially those who were female, and those raised in Mexico – more closely associated health with spirituality, as traditional Mexican medicine, or *curanderismo*, is directly connected with ritual and a more holistic sense of well-being. It is interesting to note that the few Hispanic participants who did not hold a comprehensive view of health were second generation Mexican Americans, and consequently, may have assumed Non-Hispanic cultural customs. At this point, the extent to which differing conceptions of health (as well as other culturally-based interpretive differences) may impact the quality of survey data is unclear. However, it is evident that because of varying cultural interpretations (of even a single term), participants were in effect answering two distinctly different questions.

Finally, some obtained results were unrelated to language or culture. During the course of conducting the interviews, it became obvious that, for the question *"Yesterday did you eat*

any beans such as kidney beans, refried beans, chili beans, bean soup, bean salad or lentils?", participants adopted differing interpretations of the word *bean*. Some participants viewed the question as asking about legumes only, while others included any kind of bean, even green beans. Yet, at that point, interviewers could only speculate whether there was a particular group of participants using a specific interpretive pattern. It was hypothesized that older participants and perhaps less educated participants would be less inclined to view the question as asking about legume consumption. As it turned out, based on statistical (regression) analysis, the patterns of interpretation were not related to specific demographic group membership, as "green bean error" was found to be essentially random.

3 Discussion

This paper has attempted to show how various types of analyses can be conducted using coded, semi-structured cognitive interviews which examine relationships between participant characteristics, such as ethnicity, and responses to tested survey questions. Further, based on the quantification of these codes, it may be determined that certain problems are more serious than others. The nature and causes of these problems can then be further investigated by making use of the original, qualitative interviewer notes from which the codes were developed. From the current study, it does appear that non-trivial differences may exist between Hispanics and Non-Hispanics in answering common health survey questions. Some of these are due to translation, some to cultural influences; some favor Non-Hispanics, others Hispanics. Presumably, problems with the translated version can be addressed through revisiting the translation process, or avoided in the first place by applying effective translation techniques (see Harkness, van de Vijver & Mohler, 2003; McKay, Breslow, Sangster, et al., 1996). Problems that have a basic cultural origin that transcends language may require more careful consideration of whether the questions as posed apply equally to all major groups to be included in the survey, and whether some underlying assumptions must be revisited (Ainsworth, 2000). In all cases, the mixed-method approach appeared to have significant utility.

Limitations and caveats. Several limitations to the existing study, and potential weaknesses of the evaluated methodological approach, must also be acknowledged:

(1) The practice of counting up problems within a small sample of cognitive interview subjects can be misleading, as this does not necessarily create a reliable index of problems that will then occur within in a survey field environment (Willis, 2005). Some problems obtained in a single cognitive interview may be of critical importance, to the extent that

these may in turn predict serious difficulty or source of error in the field environment for an important segment of the surveyed population.

(2) Further compounding this problem, in many pretesting studies, much smaller rounds of interviews are generally conducted than were included in the current study; restrictions by the U.S. Office of Management and Budget often limit the size of an interviewing round to no more than nine subjects. In such cases it is unlikely that strict quantification of results would supply the amount of information obtained when many more interviews are conducted, and the quantifiable arm of the mixed-method approach may be woefully insufficient.

(3) Because Hispanic/Spanish language interviews were conducted in Ohio but not in Washington D.C., the current study partially confounded ethnic group membership and region in which the interview was done; as Miller (2002) has observed, cognitive interviews that are conducted in varying regions can produce somewhat different results, and lead to varied conclusions concerning the adequacy of individual survey questions. Hence, Hispanic versus Non-Hispanic differences observed conceivably could, to some extent, reflected regional rather than ethnic or language variation[1].

(4) By far, the greatest drawback experienced in the current study was the amount of time required to ensure that all codes were applied consistently across all interviews, and to develop and clean an analyzable data set. The current approach cannot practically be implemented within cognitive testing projects that must be completed quickly or with little cost.

On the other hand, it was found that use of a mixed method approach was advantageous in providing multiple types of findings that are not entirely obtainable with traditional analytic methods for assessing cognitive interviews. Because both quantitative and qualitative analyses were used, not only could a rough measure of problem intensity for each tested question be obtained, but also the nature of potential response errors, and whether the problem appeared to be systematically related to demographic or other measured characteristics. Unlike behavior coding (Fowler & Cannell, 1996), mixed-method coding is intensively contextual in nature and allows for a more in-depth understanding of problem origins. Perhaps most importantly, this approach provides another avenue for theory building within the field of question design. For example, examining questions that are influenced by demographic characteristics will provide a better understanding of the

1 On the other hand, note that demographic variables that are commonly associated with regional differences, such as educational level and income, were fairly well controlled in the current study, so there is no particular reason to suspect a region effect in this case.

relationship between respondents' social location and response error and, in turn, will provide insight into the quality of survey estimates for particular socio-cultural groups. Currently, a multi-agency effort (involving NCHS, the U.S. Census Bureau, Bureau of Labor Statistics, and National Institutes of Health) is underway to apply such a scheme routinely to the outcomes of cognitive interviewing projects, in order to produce a dataset useful for establishing general relationships between question characteristics, respondent characteristics, and sources of error in survey questions (Miller, Canfield, Beatty, et al., 2003).

References

Ainsworth, B. E. 2000. "Issues in the assessment of physical activity in women." *Research Quarterly for Exercise and Sport* 71:37-42.

Conrad, F., J. Blair, and E. Tracy. 2000. *Verbal reports are data! A theoretical approach to cognitive interviews.* Office of Management and Budget: Proceedings of the 1999 Federal Committee on Statistical Methodology Research Conference, 317-326.

DeMaio, T. J., and J. M. Rothgeb, 1996. "Cognitive interviewing techniques: In the lab and in the field." Pp. 177-195 in *Answering questions: Methodology for determining cognitive and communicative processes in survey research,* edited by Schwarz, N., and S. Sudman. San Francisco: Jossey-Bass.

Denzin, N., and Y. Lincoln. 2000. *The Handbook of Qualitative Methods.* Thousand Oaks: Sage Publications.

Foddy, W. 1998. "An empirical evaluation of in-depth probes used to pretest survey questions." *Sociological Methods and Research* 27:103-133.

Fowler, F. J., & C. F. Cannell. 1996. "Using behavioral coding to identify problems with survey questions." Pp. 15-36 in *Answering questions: Methodology for determining cognitive and communicative processes in survey research,* edited by Schwarz, N., and S. Sudman. San Francisco: Jossey-Bass.

Gerber, E. 1999. "The view from anthropology: Ethnography and the cognitive interview." Pp. 217-234 in *Cognition and Survey Research,* edited by Sirken, M. G. et al. New York: Wiley.

Harkness, J. A., F. J. R. van de Vijver, and P. Mohler. 2003. *Cross-Cultural Survey Methods. Hoboken.* New York: Wiley.

Johnson, T. P. 1998. "Approaches to equivalence in cross-cultural and cross-national survey research." *ZUMA-Nachrichten Spezial* 3:1-39.

McKay, R. B., M. J. Breslow, R. L. Sangster, S. M. Gabbard, R. W. Reynolds, J. M. Nakamoto, and J. Tarnai. 1996. "Translating survey questionnaires: Lessons learned." *New Directions for Evaluation* 70:93-105.

Miller, K. 2002. Paper presented at the Annual Meeting of the American Association for Public Opinion Research, St Pete Beach, FL. *The role of social location in question response: Rural poor experience answering general health questions.*

Miller, K., B. Canfield, P. Beatty, K. Whitaker, and B. Wilson. 2003. Paper presented at the 2003 Federal Committee on Statistical Methodology Research Conference, Arlington, VA. *Q-BANK: Development and implementation of an evaluated-question database.*

Redline, C., R. Smiley, M. Lee, T. DeMaio, and D. Dillman. 1998. *Beyond concurrent interviews: An evaluation of cognitive interviewing techniques for self-administered questionnaires* (Working Papers in Survey Methodology No. 98/06). Washington, DC: U.S. Census Bureau. Retrieved February 12, 2004 from www.census.gov/srd/papers/pdf/sm98–06.pdf.

Snijkers, G. 2002. *Cognitive laboratory experiences. On pretesting computerised questionnaires and data quality.* Unpublished Doctoral Dissertation, Utrecht University, Netherlands.

Snijkers, G. 2004. *Cognitive laboratory experiences and beyond: Some ideas for future research.* Proceedings of the 4th Conference on Questionnaire Evaluation Standards, Zentrum fur Umfragen, Methoden und Analysen (ZUMA), 190-203.

Tashakkori, A., and C. Teddlie. 1998. *Mixed Methodology: Combining Qualitative and Quantitative Approaches.* Thousand Oaks: Sage Publications.

Tourangeau, R. 1984. "Cognitive science and survey methods: A cognitive perspective." Pp. 73-100 in *Cognitive aspects of survey design: Building a bridge between disciplines,* edited by Jabine, T. et al. Washington, DC: National Academy Press.

Tucker, C. 1997. "Measurement issues surrounding the use of cognitive methods in survey research." *Bulletin de Methodologie Sociologique* 55:67-92.

Warnecke, R. B., T. P. Johnson, N. Chavez, S. Sudman, D. P. O'Rourke, L. Lacey, et al. 1997. "Improving question wording in surveys of culturally diverse populations." *Annals of Epidemiology,* 7(5):334–342.

Willis, G. 2005. *Cognitive Interviewing: A Tool for Improving Questionnaire Design.* Thousand Oaks: Sage.

PART II: DIFFERENT ISSUES OF COMPARABILITY OR "EQUIVALENCE"

TEXT AND CONTEXT: CHALLENGES TO COMPARABILITY IN SURVEY QUESTIONS

MICHAEL BRAUN & JANET A. HARKNESS

In cross-lingual, cross-cultural studies, appropriate language and translations often play a key role in securing cross-cultural comparability. Sometimes language-as-language seems to be the foremost issue, sometimes more a blend of language, language use, and culture. At other times, problems of "equivalence" have less to do with language and how it is used than how cultural contexts frame respondents' understanding of questions. In the following we set out to explain how – against the backdrop of specific cultural contexts – cognitive and communicative processes can trigger different interpretations of survey items.

The paper provides illustrations of each of these challenges to "equivalence" and demonstrates how cultural contexts – whether linked to language or not – are related to respondents' perceptions of what questions mean in ways similar to effects noted in cognitive research on the influence of question context and co-text on respondents' perception of meaning. In fact, we suggest that question contexts are always culturally anchored contexts. The effects of a language system, its usage, and the specific cultural context are frequently difficult to disentangle.

The paper explores the relationship between language-anchored features and non-linguistic aspects of survey questions in contexts that create problems for "equivalence" or comparability. Selected examples illustrate the usefulness of an integrated framework in trying to come to terms with social science research across different cultures and languages.

1 Introduction

In cross-lingual, cross-cultural studies, appropriate language and translations often play a key role in securing cross-cultural comparability. Sometimes language-as-language seems to be the foremost issue, sometimes more a blend of language, language use, and culture. At other times, problems of "equivalence" are more problems of how the respondents' understanding

of the items is framed by their cultural context than problems of language or how language is used. Against the backdrop of specific cultural contexts, the paper illustrates how cognitive and communicative processes can trigger different interpretations of survey items.

The processes of interpretation and understanding of questions work in a very general fashion, in both national and cross-nationally comparative studies. We begin below by first discussing effects of cultural contexts, irrespective of whether a questionnaire was translated or not. We then apply these to translation, considering how source question meaning and translated question meaning can be ascertained and kept comparable.

2 Cultural Context Effects Irrespective of Translation

2.1 The role of cultural contexts

Statements, questions, and other utterances have what is generally called semantic meaning, that is, meaning(s) language users regularly associate with the words and the arrangement of words present in any given utterance. They also have pragmatic meaning, that is, meaning that is determined by the interdependence of what is said with the context in which it is said. Pragmatic considerations have an impact on how words and utterances are understood in a given context. These considerations are based on the "common ground" (Clark & Schober, 1992) which participants in the communication share, including their shared knowledge of the world. In everyday life, communication is facilitated by appropriate behavior by both the senders and the receivers. The senders of a message tailor what they say to provide others with the information that they need in order to understand the message in the way intended. The receivers contribute to the success of communication by assuming that what is said is based on common ground and by using "grounding procedures" (Schober, 1999) to verify the adequacy of their interpretation, if necessary.

Unfortunately, the situation in surveys is different from everyday communication. In standardised interviews, for instance, communication is asymmetric; interviewers are supposed to ask, respondents are supposed to answer, but not to ask. If respondents do ask, e.g. about the meaning of questions, interviewers are supposed to ensure that their answers have minimal effects on the respondents' behavior in the interview. In order to further this aim, they are recommended to respond "Whatever it means to you" (Fowler, 1992: 219), a response definitely unacceptable in everyday communication. While formalised interviewer scripts might actually reduce the impact of variations in the conduct of interviewers on the respondents' behavior, they create new problems. This is partly due to the fact that interviewer effects represent only one source of error. The wording of individual questions as well as the question order and the response category design may also violate the cooperation principle (Grice, 1975), which guides successful communication.

Respondents in surveys have a number of tasks to complete: interpret a question, generate an opinion, match the opinion to a response category ("formatting"), and edit the response taking differential social desirability of answer categories into consideration (Strack & Martin, 1987; Tourangeau & Rasinski, 1988). In order to complete all these tasks, respondents have to understand the meaning, and in the case of unclear or ambiguous questions, they try to derive the meaning from the context. The textual context of a question is represented by the rest of the questionnaire, in particular by the parts already processed: introductory texts, similar questions, the sequence of questions, and answer scales. While the question context holds much sway in social-cognition explanations of respondent behavior (see Schwarz, 1996; Sudman, Bradburn & Schwarz, 1996 for good overviews), another group of variables are less frequently used in these explanations: the personal experiences of respondents. Socio-demographic characteristics, previous behavior of the respondent, psychological or physical states, and external conditions are all related to or are even indicators of these experiences. These variables also form an important context, relevant for the respondents' behavior, such as the interpretation of questions, which has similar effects as the question context.

In addition to the contexts formed by other components of the questionnaire and the personal experiences of respondents, cultural contexts are of utmost importance in comparative research. Cultural norms, values, and experiences influence the processing of the different tasks respondents have to fulfill in an interview (Johnson et al., 1997). All three kinds of contexts provide respondents with information that can have an effect on the interpretation of a question. These contexts often operate in interaction. The interaction between question and personal-experience contexts results in the conditional context effects discussed in the social-cognition literature (Smith, 1992).

Cross-national surveys are likely to contain questions which do not resonate with or match the societal reality and the issues of public debate in some of the participating countries. Nevertheless, following the cooperation principle, respondents will assume that all the questions in a survey should make sense for them. Clark & Schober (1992: 28) suggest their reasoning may be as follows: "If the surveyor thinks this word has an obvious meaning, then it must be the meaning that is obvious to me at the moment". Whenever a salient relationship between a question and the situation in a given country can be established, respondents can be expected to establish it. Depending on the details of their cultural context, respondents can thus also be expected to perceive one and the same question differently. Researchers would be well advised to consider this when interpreting responses.

Question context effects arise when, in parsing a question, respondents process elements of the question context that were not intended as contributions to the interpretation process. Question contexts, in the narrow sense of the term "context", are provided by information which is not permanently available to respondents. Cultural context effects, on the other hand, are not the result of where a question is placed in a questionnaire but how the cultural context contributes to the way respondents process questions. Cultural context effects are based on cross-cultural differences in the saliency of different concepts and in the permanent accessibility of pertinent information. The everyday reality and the frequency of particular events in a society lead to the formation and stabilization of schematic structures.

2.2 Cultural context effects

Different interpretations of a question that are linked to cultural factors can be discussed in terms of what are sometimes called "framing effects". These result from differences in framing conditions. Following Stocké (2002), we distinguish between framing effects related to question ambiguity, heuristic considerations, and schemas relevant for a given topic or question. Respondents naturally parse questions in terms of their knowledge of the world and understanding of the interview situation, in other words, they use their cultural knowledge to help them interpret question meaning.

An item from the 1994 ISSP module on family and gender (Braun, 1994; Zentralarchiv 1994) illustrates how this interpretive process may work. The item "A pre-school child is likely to suffer if his or her mother works" leaves several informational components unspecified, including the age of the child, the amount of labour-force participation respondents are to assume for the mother in question, and other considerations such as whether the father is employed outside the home or another adult relative is available to care for the child (Braun, 1998, 2003). Ambiguity-based framing effects will result: respondents will assemble the scenario culturally most salient for them, filling the "gaps" by activating schemata linked to the social realities of their given society. Relevant details might include the rate of female participation in the labor-force, respondents' knowledge about the availability of part-time jobs and whether crèches are common for very young children. Thus, in some countries respondents might tend to picture the child mentioned in the question as an infant, knowing that older children could be cared for by other means, whereas in other contexts where child care is generally difficult or where mothers regularly stay at home until children are past the toddler stage, 5-year olds might come to mind. In some societies, given the facts of labor-force participation, respondents will assume full-time employment of both parents or scarcity of paid jobs all-round, irrespective of gender.

Using a set of related questions, Braun (2003) was able to demonstrate large differences in responses and understanding in an experiment carried out in 1998/99 in eastern and western parts of Germany (geographically the former East and West Germanys). Respondents were asked to evaluate the suffering of a 3-year old child in different situations (e.g. when both parents work full-time, when the mother works full-time, but the father only part-time, etc.). Respondents who have the same values on the ISSP item above which does not provide explicit detail of the schema involved, differed greatly in their evaluation of the effects on a 3-year old child. Former East Germany had and still has a much higher incidence of childcare facilities for young and very young children and a historically much higher participation rate of women in the labour force. Thus, for former East Germans, if they draw on the East German history of child care and labour force participation, the item has a higher item difficulty.

Questions that are perceived by respondents as not pertinent to an ongoing public debate, or that are not seen to address what respondents consider to be the relevant aspects of a topic, are also problematic. Feeling the question is in some sense irrelevant, respondents might not be motivated to process all the information presented in the question. Instead, they could focus on individual features that do link up positively or negatively with their own views on the topic and take that as a basis for answering the question. Heuristic-based framing effects will result. For example, the information that the woman is to be thought of as working full time in the item "All in all, family life suffers when the woman has a full-time job" might be ignored by respondents who want to demonstrate that they have non-traditional attitudes but, at the same time, cannot imagine that both parents work full-time because they live in a society where it is difficult to juggle work and the family.

Finally, schema-based framing effects as the consequence of an automatic activation of mental structures can result when items holistically activate a schema, over and above the literal meaning conveyed by the question text. The ISSP item "It is not good if the man stays at home and cares for the children and the woman goes out to work" provides an illustration of this. The question is intended to test whether respondents are in favour of a reversal of an arrangement by which men go out to work and women stay at home and care for the children. This arrangement, however, also presupposes that it is accepted that some person should stay at home to care for the children. In other words, it implicitly adopts an ideology schema (cf. Fiske & Taylor, 1991 on ideology schema) that presents labour-force participation of both parents as incompatible with looking after children well. In societies in which this schema does not exist, respondents will be at a loss. Respondents in former communist countries are not likely to read the item the way it was intended. They might see it as implying the man does not have a job and is therefore at home. Much would argue against endorsing an item that is understood to refer to men

being at home because they are out of work. The intended notion of a role reversal will also not make much sense in regions of eastern Germany in which participation of women in the labor force used to be compulsory and is now welcomed for economic reasons.

Cultural contexts and cultural context effects pose problems for all cross-cultural comparative social research, irrespective whether translation is involved or not. In some of the examples mentioned above, the questions were identical in one language (German). It was the cultural background respondents brought to the questions that differed.

3 Questionnaire Translation and Questionnaire Meaning

We now turn briefly to consider the implications of the above for the translating process and for translated questionnaires. Questionnaires are usually designed in one language, often English, and then translated into the other languages required in order to interview populations that cannot be interviewed in the language available. In translation jargon we speak of translating out of a "source" language into a "target" language. Questionnaires are translated for three main purposes – for cross-national survey projects, for within-country research in countries with several official languages, and for projects in which it is necessary to include populations that do not speak the majority language of a given country. In each case, we can expect the cultural context to vary across populations.

The technicalities of team translation procedures have been discussed elsewhere (Harkness, 2002, 2003, 2004), the kinds of specific problems that arise (Harkness, 2003; Harkness et al., 2004), as well as issues of adaptation and cultural tailoring (Harkness, 2004) and we do not discuss these here. Instead, we focus on meaning in the context of translation, given that meaning is co-constructed and depends on semantics, immediate context and co-text, discourse conventions and pragmatics, and, as just outlined, the socio-cultural framework in which respondents are embedded.

3.1 Questions, words, and meaning

We illustrated earlier how respondents use knowledge of the(ir) world in interpreting what questions are intended to mean. Words and combinations of words are usually open to multiple interpretations, depending on the words themselves, the context of utterance, the participants involved, the foregoing communication (co-text), the common ground shared by participants, and other pragmatic considerations related to how we communicate and the role that pragmatic factors of various kinds play in that. Space restrictions prevent us from discussing these here.

In sum, question meaning is not determined by the wording of questions alone. The distinction between what an utterance or part of an utterance is intended to mean (often called *speaker intended meaning*) and what recipients of an utterance understand it to mean (called *perceived meaning*) is an important one in unravelling how meaning is negotiated in communication and is also frequently open to misunderstanding.

Chafe's (1980) famous pear stories, stories generated on the basis of visually presented stimuli, are an excellent illustration of the fact that different cultures perceive the same material differently. Tanzer (2005) discusses cultural effects in visual material used in educational tests. As illustrated earlier, differences in cultural framing may mean that a reading of a question salient for one cultural group is not the reading that is salient for a second cultural group. In other words, perceived meaning may well differ from cultural group to cultural group. Within country research has shown that cultural groups that form part of a single larger society also differ in how they understand and respond to questions (e.g., Johnson et al., forthcoming; Miller, 2003; Willis, 2004). We can expect differences will increase as the distance between cultural, societal, language and pragmatic systems increases.

Whenever translation is called for in survey research, differences in these systems will exist across the survey populations interviewed. As a result, even if a translation may be judged to be technically adequate, the interpretation a question receives can easily differ across different cultural groups.

3.2 Researcher expectations for translations

In producing translations of survey questions, researchers intend to ask the same questions in different contexts and languages. Few would expect survey translators to produce word-for-word translations in order to do this and it would be an odd translator who actually managed to work on this basis. This said, survey translators are generally expected to convey the semantic content of questions faithfully and to stay as close to the original as possible. This is commonly taken to mean that translators should try to convey also finer details considered relevant in terms of measurement or question design. Thus if a question includes the phrase "if any/if at all" (e.g., *To what extent, if at all, do you ...*), which is a survey strategy to accommodate respondents for whom the question may not apply, translators would normally be expected to match this in translation. As noted elsewhere, including such details may produce awkward or more complicated questions (Harkness et al., 2004). Alternatively, if the source questionnaire refers to "foster children" or "god children", the target questionnaire might also be expected to refer to these. This, as it turns out, is also not an easy matter; societies differ linguistically and legally in the distinctions they make

between various kind of "children" and forms of caring for the children. Then again, if the source question formulation is symmetrically organised and asks "to what extent do you *agree or disagree* with the following statements", translators will be expected to convey this deliberately balanced phrasing, rather than producing something more like "Do you agree with the following statements?". However, not all languages can match the agree/disagree pair. In addition, it has been suggested in various places that Hispanic populations will prefer to indicate agreement or the opposite rather to distinguish degrees of agreement or disagreement. The same applies to the translation of single key words or phrases. Thus if the source question asks about pride "in the arts and literature", as in the 1995 ISSP module on National Identity, translators might be expected to find terms that cover exactly everything covered by "arts and literature" in English, without changing stimulus or respondent burden. This again can prove to be a challenge. In order to include the performing arts in German, for example, an additional and rather formal descriptor would be necessary.

Thus translation in the form often expected for surveys can best be likened to a balancing act, requiring know-how, practice, talent but also the confidence on the part of the translator to speak up when close translation will not work. Translators working into languages and cultures at a great distance from the source questionnaire language and culture have, in any case, less opportunity to stick close to the source text if they want to avoid the text sounding really awkward or nonsensical. An overview of problems related to close translation can be found in Harkness et al. (2004).

The European Social Survey tries to encourage participating countries to seek functional equivalence of stimulus, partnered with comparability of semantic content. In other words, if a source question contains the phrase "race or ethnicity" and one or both of these terms cannot be directly translated for any of a variety of reasons, translators have the leeway to seek a phrase that can be used in their cultural context to secure information about cultural and/or genetic ancestry. At the same time, this freedom (and responsibility) is a challenge for survey researchers. Readers who speak Spanish, German or French, for example, might find it instructive to compare the different approaches to translation evidenced in translations in the same language in recent ISSP and ESS studies. Questionnaires for both surveys can be downloaded free of charge from the web.

Briefing sessions for translators should make clear what in a given project "stay close to the original" means. Otherwise, translators might focus on words rather than on the intended meaning of questions. Research on students of translation shows that inexperienced translators work more on the level of words than on the level of unit meaning, thereby increasing the likelihood of too close and ineffective translation (Kussmaul,

1995). Unfortunately, survey research often employs inexperienced people who may translate infrequently and have little professional training.

Thus, as we understand survey translation, the task to be undertaken goes considerably beyond commissioning translators to produce a literal or semantically faithful translation. This is not to say that survey translations should be free translations. Measurement issues in questions often require careful retention of scope or emphasis across translations, for example.

Like respondents reading a source question, translators scan questions for sense or intended meaning on the basis of their knowledge of the context, the co-text (questions surrounding the question to be translated), and the socio-cultural framework within which a given question is embedded. Unlike respondents, they are trained to parse not only to understand but to parse with the specific aim of translating. Their training and skills help them identify potential ambiguities and translation problems. This is one reason why translators can be very useful proof-readers for draft questionnaires and can be helpful in developing questionnaires intended for comparative use.

At the same time, without understanding the measurement properties of questions and answer categories, it may be difficult for translators to be sure what a question is really intended to do. This is one of the reasons why team procedures are suggested for survey translation. By setting up a team, people with the necessary language and translatory skills can be brought together with people who understand the goals and structure of an instrument (cf. Harkness & Schoua-Glusberg, 1998; Harkness, 2003; Harkness et al., 2004). Thus translators who have an understanding of survey question design are invaluable. Other things being equal, the better briefing that translators can be given, the better equipped they will be. A technical understanding of specific design components in questions can guide their appraisal of the source question and their decisions on how best to translate this.

3.3 Using translation to inform questionnaire design

It is sometimes assumed that questions that performed well in one or more contexts will perform well in other contexts too. Since tried and tested questions may also be adopted into a survey without pre-testing for the new context, problems they may have for a new socio-cultural and linguistic setting may only become apparent after the event. The literature abounds with examples of questions that have "gone wrong". Harkness (1995, 2003, 2004) has suggested using rough or "advance" translation to inform source questionnaire design for comparative projects. The idea is that translators embedded in the socio-cultural context for which they are translating will be able not only to point to linguistic

challenges in the source question with regard to translation but also to pragmatic and socio-cultural issues. In our experience, only considering the questions in their English form disguises some of the problems lying in wait for implementation in other cultures. Optimal protocols for such procedures remain to be developed. If researchers alone are involved in producing these first draft translations, we run the risk that they, like inexperienced translators, would be influenced by the source questions and could fail to note problems. If we use only translators without knowledge of survey design, we run the risk that they do not perceive the measurement issues at stake.

This notwithstanding, the automatic parsing for meaning that is part of the professional translator's training is invaluable for developing survey questions intended for multiple socio-cultural contexts. These rough translations are an early alarm-signal strategy. They are not intended as a substitute for testing at the much later stage of final draft questionnaires. Finally we note that different populations sometimes share a language, that is, they use a different regional form of the language (American, British or Australian English). While the linguistic differences may not be great, the socio-cultural differences may result in differences of interpretation. Care should be taken to adjust wording if necessary.

4 Conclusion

Poor translations of good questions mean respondents read and respond to a question they should not have been asked. Researchers lose the opportunity to ask the questions they intended. However, technically well-translated questions that are understood differently in different cultures are equally problematic. In order to make sure that respondents anchored in different cultural contexts perceive one and the same intended meaning, procedures of question development, question testing, and question translation need to change.

By providing detailed specification of the measurement intended and the scenario envisaged in the source questions, translators and adaptors could produce questions that allow respondents to focus on the intended meaning and the appropriate scenarios in other contexts. In saying this, we obviously envisage that translation be understood as translation of intended meaning and intended measurement goals and not simply translation of semantic content. A major goal for future research must be to develop a systematic scheme to identify what aspects of questionnaires should be given particular scrutiny for comparative design, with or without language differences.

References

Braun, M. 1994. "The International Social Survey Program (ISSP)." Pp. 305-311 in *Social Statistics and Social Reporting in and for Europe,* edited by Flora, P. et al. Bonn: Informationszentrum Sozialwissenschaften.

Braun, M. 1998. "Gender roles." Pp. 111-134 in *Comparative Politics. The Problem of Equivalence,* edited by van Deth, J. W. London/New York: Routledge.

Braun, M. 2003. "Communication and social cognition." Pp. 57-67 in *Cross-Cultural Survey Methods,* edited by Harkness, J., F. J. R. van de Vijver, and P. Ph. Mohler. Hoboken NJ: Wiley.

Brislin, R. W. 1970. "Back-translation for cross-cultural research." *Journal of Cross-Cultural Psychology* 1:185-216.

Brislin, R. W. 1990. "Applied cross-cultural psychology. An introduction." Pp. 20-22 in *Applied Cross-Cultural Psychology,* edited by Brislin, R. W. Newbury Park: Sage.

Chafe, W. L. 1980. *The Pear Stories. Cognitive, Cultural and Linguistic Aspects of Narrative*. Norwood NJ: Ablex.

Clark, H. H., and M. F. Schober 1992. "Asking questions and influencing answers." Pp. 15-48 in *Questions about Questions: Inquiries into the Cognitive Bases of Surveys,* edited by Tanur, J. M. New York: Russell Sage.

Fiske, S. T., and S. E. Taylor 1991. *Social Cognition.* Second Edition. New York: McGraw-Hill.

Fowler, F. J. Jr. 1992. "How unclear terms affect survey data." *Public Opinion Quarterly* 56:218-231.

Grice, H. P. 1975. "Logic and conversation." Pp. 41-58 in *Syntax and Semantics 3, Speech Acts,* edited by Cole, P., and J. L. Morgan. New York: Academic Press.

Harkness, J. 2002. *Translation: Final Report on Round One to the European Commission.* ESS. http://europeansocialsurvey.org/

Harkness, J. 2003. "Questionnaire Translation." Pp. 35-56 in *Cross-Cultural Survey Methods,* edited by Harkness, J., F. J. R. van de Vijver, and P. Ph. Mohler. Hoboken NJ: Wiley.

Harkness, J. 2004. *Supplement on Translation Procedures and Requirements for Round Two of the ESS.* http://europeansocialsurvey.org/

Harkness, J. 2004. *Instrument Translation, Adaptation and Cultural Tailoring.* Invited Symposium presentation. 28th International Congress of Psychology. Beijing, China, August 8 – 10.

Harkness, J. 2004. "Overview of Problems in Establishing Conceptually Equivalent Health Definitions across Multiple Cultural Groups." Pp. 85-90 in *Eighth Conference on Health Survey Research Methods,* edited by Cohen, S. B., and J. M. Lepkowski. Hyattsville: US Department of Health and Human Services.

Harkness, J., B. Pennell, and A. Schoua-Glusberg. 2004. "Questionnaire Translation and Assessment." In *Methods for Testing and Evaluating Survey Questionnaires,* edited by Presser, S., J. Rothgeb, M. Couper, J. Lessler, E. Martin, and E. Singer. New Jersey: John Wiley and Sons.

Harkness, J. A., and A. Schoua-Glusberg 1998. "Questionnaires in translation." *ZUMA-Nachrichten Spezial* 3:87-127.

Johnson, T. P., Y. I. Cho, A. Holbrook, D. O'Rourke, R. Warnecke, and N. Chávez (forthcoming). "Cultural variability in the comprehension of health survey questions." *Annals of Epidemiology.*

Johnson, T., D. O'Rourke, N. Chavez, S. Sudman, R. Warnecke, L. Lacey, and J. Horm. 1997. "Social cognition and responses to survey questions among culturally diverse populations." Pp. 87-114 in *Survey Measurement and Process Quality,* edited by Lyberg, L., P. Biemer, M. Collins, E. de Leeuw, C. Dippo, N. Schwarz, and D. Trewin. New York: Wiley.

Kussmaul, P. 1995. *Training the Translator.* London: John Benjamins.

Miller, K. 2003. "Conducting Cognitive Interviews to Understand Question Response Limitations." *American Journal of Health Behavior* 27 (Suppl 3):264-272.

Schober, M. F. 1999. "Making sense of questions: an interactional approach." in *Cognition and Survey Research,* edited by Sirken, M. G., D. J. Herrmann, S. Schechter, N. Schwarz, J. M. Tanur, and R. Tourangeau. New York: Wiley.

Schwarz, N. 1996. *Cognition and Communication: Judgmental Biases, Research Methods, and the Logic of Conversation.* Mahwah NJ: Erlbaum.

Smith, T. W. 1992. "Thoughts on the nature of context effects." Pp. 163-186 in *Context Effects in Social and Psychological Research,* edited by Schwarz, N., and S. Sudman. New York: Springer.

Sperber, D., and D. Wilson. 1986. *Relevance: Communication and Cognition.* Oxford: Basil Blackwell.

Stocké, V. 2002. *Framing und Rationalität. Die Bedeutung der Informationsdarstellung für das Entscheidungsverhalten.* München: Oldenbourg.

Strack, F., and L. L. Martin 1987. "Thinking, judging, and communicating: a process account of context effects in attitude surveys." Pp. 123-148 in *Social Information Processing and Survey Methodology,* edited by Hippler, H.-J., N. Schwarz, and S. Sudman. New York: Springer.

Sudman, S., N. M. Bradburn, and N. Schwarz 1996. *Thinking about Answers. The Application of Cognitive Processes to Survey Methodology.* San Francisco: Jossey-Bass.

Tanzer N. K. 2005. "Developing Tests for Use in Multiple Languages and Cultures: A Plea for Simultaneous Development." Pp. 235-263 in *Adapting Educational and Psychological Test for Cross-Cultural Assessment,* edited by Hambleton, R. K., P. F. Merenda, and C. D. Spielberger. Hillsdale, NJ: Erlbaum.

Tourangeau, R., and K. A. Rasinski. 1988. "Cognitive processes underlying context effects in attitude measurement." *Psychological Bulletin* 103:299-314.

Willis, G. 2004. "Overview of Methods for Developing Equivalent Measures Across Multiple Cultural Groups." Pp. 91-96 in *Proceedings of the Eighth Conference on Health Survey Research Methods (DHHS No. PHS-04-1013),* edited by Cohen, S., and J. Lepkowski. Hyattsville, MD: National Center for Health Statistics.

Zentralarchiv für Empirische Sozialforschung. 1997. *ISSP 1994 - Family and Changing Gender Roles II. Codebook ZA-No. 2620.* Köln: Zentralarchiv.

MEASURING ATTITUDES TOWARDS IMMIGRATION ACROSS COUNTRIES WITH THE ESS

Potential Problems of Equivalence

NINA ROTHER

This paper analyses the functional equivalence of attitudes towards immigration in internationally comparative research. In order to provide a knowledge base on attitudes towards immigration in Europe, a special module was included in the ESS round 1. However, attitudes towards immigration are not easy to measure and especially difficult to compare across cultures as the underlying concepts of immigration and immigrants may vary across countries.

Therefore a cross-cultural comparison of attitudes towards immigration can only be done when functional equivalence is given. Functional equivalence of data can be affected by three kinds of bias: construct bias, method bias and item bias. In order to analyse the functional equivalence of ESS immigration items, a secondary analysis of the ESS data was conducted.

The results show that some of the ESS measures might not be regarded as functionally equivalent. Some possible solutions are outlined.

1 Introduction

To know about attitudes towards immigration is essential, not only for researchers, but especially for politicians. Immigration plays a more and more important role in European societies nowadays, as can be seen in growing concerns about refugee issues or the importance to lower restrictions for highly skilled migrants that are needed in certain industries. If we just think of recent events as the EU enlargement or the Cap Anamur incident, followed by discussions about refugee camps in North Africa, it is getting clear how important it has become again to know more about what "attitudes towards immigration" are like in European states.

In order to provide researchers with knowledge and a database on attitudes towards immigration within Europe, the European Social Survey (ESS) round 1 contains a special module on attitudes towards immigration. So the ESS seems to be ideal for answering those questions on what those attitudes are like in Europe. However, it is not easy to measure, and eve more difficult to compare attitudes towards immigration across cultures. Different migration histories and policies in the different European countries make it hard to think of a common understanding of what immigration and immigrants are. Therefore, before starting with substantive analyses, it should always be proved that the data are comparable, i.e. that they are functionally equivalent.

After a short discussion of functional equivalence, types of biases and possible detection methods, a few potential biases found within the ESS data on attitudes towards immigration will be shown on an exemplary basis.

2 Bias and Equivalence

Lots of bias definitions and huge error lists from different research areas such as cross-cultural survey research (Braun, 2003) or cross-cultural psychology can be found (Johnson, 1998; Hui & Triandis, 1985; Berry et al., 1993; van de Vijver, 1998). For the purpose of this paper, it is especially relevant to focus on the psychological perspective of equivalence. For cross-cultural psychologists, data are equivalent if the underlying theoretical concepts, as well as the scores, have a similar meaning in the different countries. The negative counterpart of equivalence – bias – is used as a generic term to describe all nuisance factors that threaten the validity of cross-cultural comparisons (van de Vijver & Leung, 1997). Following the classification of van de Vijver & Tanzer (1997), cross-cultural psychologists further distinguish between three levels of bias: construct, method and item bias.

We are talking of *construct bias* if a not identical construct is measured across cultural groups. van de Vijver & Poortinga (1997) give an overview of possible causes of construct bias. They mention, for example, an incomplete overlap of definitions or an incomplete coverage of the construct – which is also called construct underrepresentation (Embretson, 1983). Also, a poor sampling of all relevant behaviours or a differential appropriateness of the questionnaire content can cause a construct bias. In the present case of attitudes towards immigration in the ESS, a construct bias due to an incomplete overlap of definitions could be the case if the questionnaire does not specify the term “people who come to live here” so that respondents in some countries would include refugees while respondents in other countries would not include them because of a different composition of immigrants in those countries.

In contrast to method and item bias, a construct bias affects the most basic level of data: the *structural* comparability of data, which comes first, i.e. before the comparison of the level of attitudes. So it always has to be analysed first if structural equivalence is given. Only then, level-oriented analyses are reasonable. Such level-oriented analyses can then still be affected by method and item bias.

The standard way to detect a "construct bias" is to compute factor analyses and then to look for differences in the factor solutions (van de Vijver, 2003). Differences in the factor solutions of different countries then indicate a construct bias. Two different kinds of factor analyses allow finding out about construct bias: On the one hand a confirmatory factor analysis for multi-groups can be computed (see for example Bollen, 1989), which allows various tests of model fit but does not offer an index of factorial agreement.

On the other hand an exploratory factor analysis using *target rotation*, also know as Procrustes Rotation, can also help to find out if structural equivalence is given (Harman, 1976). A factor analysis has to be computed in which the country factor loadings are rotated towards the general solution in order to correct for the arbitrariness of rotations in factor analyses. After that, as an indicator of factorial agreement, different coefficients of agreement like the identity coefficient, the additivity coefficient or the proportionality coefficient, which is also known as *Tucker's phi* can be calculated (van de Vijver & Leung, 1997). As the latter coefficient is insensitive to multiplications but influenced by additions, it seems to be the appropriate coefficient for comparisons of the structural equivalence of the immigration dataset.

But apart from factor analysis, it is always wise to collect additional data on the culture-level to investigate the underlying construct in detail, possibly by conducting cognitive interviews. It is only by getting more information on the different function of the biased items in the different countries, that you can find out what the real reasons for the construct bias have been and how the data can be corrected.

The second type of bias, *method bias*, is a bias that occurs within the scores due to particular characteristics of the instrument or the administration. It therefore affects scores at the level of the whole instrument (van de Vijver & Leung, 1997). Possible causes are many and various, such as a differential response style – like social desirability for example –, interviewer effects or communication problems between the respondent and the interviewer. The presence of a method bias can be seen through significant cross-country differences. However, a mixture of valid cross-country differences and method bias cannot be excluded so that it is wise to utilise other methods such as cognitive tests and monotrait-multimethod designs here as well.

Finally, we are talking of an *item bias* if persons from different cultures, but with the same standing on the underlying construct have a different score on the item (van de Vijver & Leung, 1997). Causes of item bias are measurement artefacts at the item level such as a poor item translation, an inadequate item formulation or differences in the appropriateness of the item content. In order to detect an item bias, a reliability analysis, for example by comparing Cronbach's α, and all exploratory methods, such as a comparison of means, Item-Response-Theory, Multidimensional Scaling or ANOVA can be used. But nevertheless, a mixture of valid cross-country differences and item bias again can occur, so that cognitive interviews to find out how the items are understood in different countries, might be useful here as well.

Although method bias definitely is important, I will now only discuss and analyse the presence of construct and item bias in the ESS data on immigration.

3 Dataset, Variables and Countries

The database for my analyses was the immigration module (D) of ESS round 1, edition 4.1. In principle, there are 58 questions on topics related to different immigration phenomena that could be analysed. However, all items on attitudes towards refugees (D49-D55) were excluded because I did not want to include attitudes towards a subgroup of immigrants but only general attitudes. All items that measure perceptions or estimations rather than attitudes were excluded, too. In detail, the items D1-D3, D38-D39, D47-D58 were excluded and 39 items on "attitudes towards immigration" remained for further analyses.

As for the countries to analyse, 20 countries will be compared and constitute the basis for all of the following analyses: Austria, Belgium, the Czech Republic, Denmark, Finland, Germany, Greece, Hungary, Italy, Ireland, Luxembourg, the Netherlands, Norway, Poland, Portugal, Slovenia, Spain, Sweden, Switzerland and the United Kingdom[1].

For the analyses of construct and item bias in this paper, however, I will focus on 5 countries which can serve as examples of the 20 countries. Germany and the UK represent the *classic immigration countries*. In these two countries, immigration has a very long history beginning with migrants from the former colonies immigrating to the UK and Polish mine workers settling in the German Ruhr basin at the beginning of the 20th century. In the early 1970ies both countries filled their labour shortages with workers mainly from Southern Europe. Nowadays a lot of refugees and asylum seekers still arrive in both countries as well as other migrants who come for reasons of family reunification. As *new immigra-*

1 For my analyses, there were 20 rather than 21 countries as I decided to exclude Israel from the beginning. Israel is simply too far away from the European context especially with regard to its immigration history.

tion countries, Italy and Portugal were chosen. Both countries have been emigration countries for a long time, sending lots of their inhabitants to Northern Europe. Only in the last few years, a back-migration has started. Additionally, both countries now have to deal with the growing problem of illegal immigrants that arrive on the shores of both countries. Finally Luxemburg, with its multicultural and multilingual society serves as a very interesting *special case*. The OECD data for 2000 (OECD, 2003) show that the percentage of foreigners in Luxembourg with about 37% lies well above the EU average. The rate of EU foreigners with almost 78% is even more impressive. Apart from this remarkable composition of the population, Luxembourg has three official languages: French, Luxembourgish and German and is therefore unique, regarding its societal structure.

4 Analysis

Before analysing the data regarding possible biases and therefore a functional inequivalence, some preliminary analyses had to be conducted.

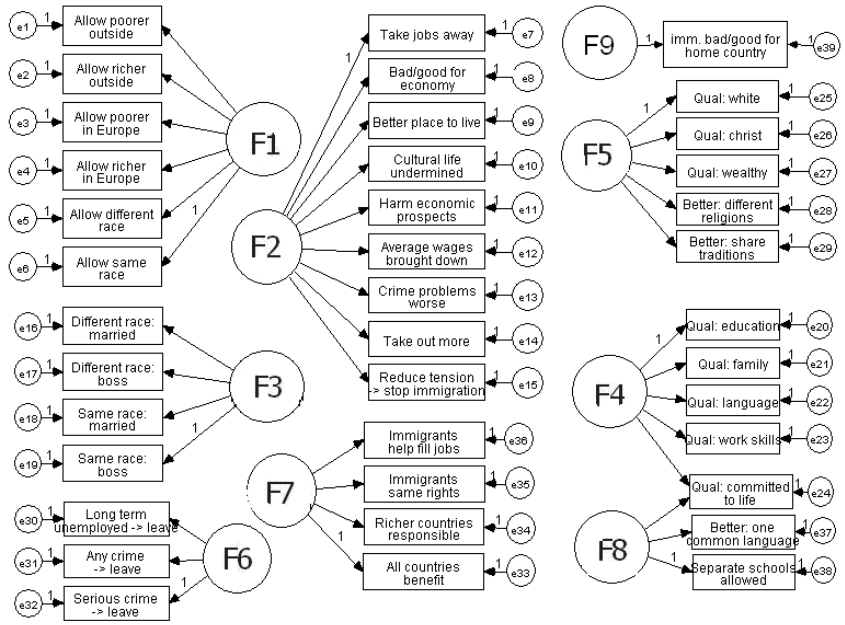
4.1 Preliminary analysis: latent structure of variables

First of all, the underlying factor structure of the 39 items had to be found out. This factor structure should assign all of the 39 items to a smaller number of factors. Furthermore, this factor structure should be a general one and therefore apply to all 20 countries in the same way.

As Christian Dustman, Ian Preston and their colleagues, who developed this immigration module for the ESS, have not yet postulated a concrete structure of their 58 items, I had to create such a structure myself by looking at the items and the theory. The result, after conducting various factor analyses and comparing the results, was a model of 9 factors. These 9 factors can be interpreted as follows:

- Factor 1 measures attitudes towards the *quantity of immigrants* that should be allowed to come into the country. It consists of six items, namely D4-D9.
- Factor 2 comprises attitudes towards the *effects of immigration on the host country*. For example, the item "immigrants take jobs away" (D25) is included here, as well as D18-D19, D26-30 and D44.
- Factor 3 captures which feelings people have about foreigners in their *personal environment*. Four items, D34-37, indicate how much one would mind or not mind if a person having either the same race or a different race would be appointed as one's boss or married a close relative.
- Factor 4 – *qualification* – indicates, how important it is that immigrants bring with them different qualifications such as a high education, a close family in the host country or the ability to speak the host country's language (D10-D12, D16).

Figure 1 Underlying Factor Structure of the 38 Immigration Items



Structural Equivalence of overall Factor Structure

- Factor 5 indicates which *cultural characteristics* like "being white" (D14), "coming from a Christian background" (D13) or "being wealthy" (D15) are important for immigrants to come. However this factor also comprises two items, which indicate how much a culturally homogenous society is desired by the respondents (D40-D41).
- Factor 6 measures *attitudes towards a deportation policy* and indicates under which circumstances people favour that immigrants are made to leave the country. This factor comprises the items D21 and D23-D24.
- Factor 7 holds a *general evaluation of immigration*, for example whether immigrants should get the same rights as everyone else (D22). Apart from this item, D20 as well as D32-33 also belong to this factor.
- Factor 8 captures attitudes towards a *cultural integration* of immigrants. Three items show high loadings on this factor, namely "better for a country if there is one common language" (D 42), "separate schools should be allowed" (D43) and "immigrants should be committed to the way of life" (D17).

- Factor 9 measures the perceived *effects of immigration on the immigrants' home country*. As it only consists of the item "immigration good/bad for those countries in the long run" (D31), this factor will be excluded from further analyses so that 8 factors remain.

Figure 1 shows the visualisation of the underlying factor structure, which could also be used for analysing it using confirmatory factor analysis.

4.2 Structural equivalence of overall factor structure

Before the data can be analysed on the substantive level by comparing the factor scores across countries, it has to be proved that the underlying factor structure applied to all of the analysed countries, to make at least sure that no construct bias has occurred. Therefore, an exploratory *principal component analysis with target rotation* is computed to find out if the same underlying factor structure can be assumed in the five selected countries. This is done by computing a principal component analysis in which the country factor loadings were rotated towards the general solution, that is the factor loadings of all 20 countries taken together. Tucker's phi was calculated afterwards as indicator of factorial agreement.

Table 1 shows the results for the 8 factors in the 5 countries. According to the suggestions of van de Vijver & Leung (1997), coefficients smaller or equal .85 have to be interpreted by all means as an indicator of missing factorial agreement. Thus factors 1 to 4 and also factor 7 show high values of factorial agreement in all countries. So it seems that the country factor structures correspond to the general assumed factor structure and no construct bias is given for these factors. In Great Britain, this also applies to all factors. Therefore it can be assumed that the British factor solution is the same as the general one and structural equivalence is given in Great Britain. This might be true because the questionnaire was developed in English, by British researchers who probably know best the situation in their home country and also have experience as to how questions are interpreted and answered in Great Britain. In Great Britain, the whole questionnaire was also pretested and possible problems thus could have been eliminated before the fieldwork started[2].

2 The second country where the items have been pretested were the Netherlands. It is a very interesting result that also in this country no deviations from the general factor structures could be found. Therefore, it seems that pretesting and adapting the item formulation according to the results of the pretests, is quite an important part of the questionnaire development, leads to big improvements of the data quality and can avoid time-consuming analyses.

Table 1 Values of Factorial Agreement Using Tucker's phi

	F1	F2	F3	F4	F5	F6	F7	F8
Germany	.99	.99	.99	.99	.98	.85	.94	.88
UK	.99	.99	.99	.99	.98	.96	.96	.96
Italy	.98	.98	.98	.94	.63	.94	.89	.82
Luxembourg	.98	.96	.96	.92	.82	.78	.91	.88
Portugal	.97	.96	.97	.95	.87	.92	.90	.53

Source: ESS round 1, edition 4.1; weighted with dweight; values equal or below .85 in italics

For the four countries other than Great Britain, the values are sub-optimal. Regarding factor 5, Italy and Luxembourg are problematic as they have quite small coefficients with a Tucker's phi of only.63 in Italy and .82 in Luxembourg. The same is true for Germany and Luxemburg for factor 6, for which Germany shows a Tucker's phi of .85 and Luxembourg of .78. Italy also has a small value of factorial agreement for factor 8 (phi = .82) as well as Portugal (phi = .53). So it has to be assumed that it has to be dealt with a construct bias in these four countries for these three factors.

A low factorial agreement in principle can have two reasons. Firstly, the loadings of the items that belong to the biased factor have evidently higher or lower loadings on this factor than in the general factor solution. The second possible reason for a low factorial agreement of a factor is that other items, which in fact do not belong to the biased factor, show too high loadings on it.

In the following analyses, it will be checked for factor 6 and factor 5 which of the two possible reasons of construct bias can be found. Furthermore, it will be checked what the reasons for these deviations could be and if they are due to an item bias of some of the items.

4.3 Structural equivalence of factor 6 – "deportation policy"

Factor 6, measuring attitudes towards a deportation policy, showed a missing factorial agreement in Germany and Luxembourg. Possible reasons for this construct bias will now be looked at, firstly by analysing the factor loading matrices, then by going down on the item level and comparing item means and finally by checking the linguistic equivalence of the factor 6 items.

Comparison of factor loadings

The factor loadings of factor 6 and other items on the biased factor 6 are given in Table 2. In the upper part of this table, you can check if the loadings of the three factor 6 items of the two biased countries show too high or too small values by comparing them with the loadings of these items of the general solution.

Table 2 Factor Loadings of Selected Items on Factor 6

		All countries	GER	LUX
F6	Serious crime → made to leave	.81	*.55*	*.51*
	Any crime → made to leave	.78	*.58*	*.66*
	Long term unemployed → made to leave	.50	.47	.43
F2	Immigration bad or good for country's economy	.07	.26	.19
	Taxes + services: immigrants take out more/less than put in	.07	*.28*	*.24*
	Immigrants make country worse / better place to live	.12	*.31*	.21
	Average wages/salaries generally brought down	.12	-.08	*-.43*
	Country's crime problems worse / better	.18	.31	*.43*
F7	Richer countries responsible to accept people	-.08	-.07	*.16*
	All countries benefit if people can move where skills needed	.06	*.40*	.18
	Immigrants help to fill jobs where there are shortages	.02	*.25*	-.01

Source: ESS round 1, edition 4.1; weighted with dweight; factor loadings higher than .2 and a deviation of more than .15 from the general loadings in italics.

It can be seen, that in both biased countries, the 3 items belonging to factor 6 have lower loadings than in the general solution, especially the items "serious crime → made to leave" and "any crime → made to leave". Therefore, it can be assumed that in Germany and Luxembourg the impact of attitudes on deportation in case of criminal behaviour on this factor is much weaker in these two countries than in the general solution. So in these two countries, factor 6 measures less attitudes towards deportation in case of criminal behaviour but rather a general evaluation of deportation of foreigners.

Another possible reason for a low factorial agreement of a factor could be, as discussed above, too high factor loadings of items that do not belong to the biased factor 6. In the lower part of Table 2, you can see that in Germany and Luxembourg variables that should only belong to factor 2 and 7 have quite high loadings on factor 6. In Germany this is true for the items "immigration bad / good for country's economy", "taxes + services: immigrants take out more / less than put in", "immigrants make country worse / better place to live", "all countries benefit if people can move where skills needed" and "Immigrants help to fill jobs where there are shortages". In Luxembourg, the items "taxes + services: immigrants take out more / less than put in", "average wages/salaries generally brought down", "country's crime problems worse / better" and "richer countries responsible to

accept people" have too high loadings on factor 6[3]. As factor 2 measures general attitudes towards the effects of immigration on the host country and factor 7 captures a general evaluation of immigration, it seems that in Germany and Luxembourg factor 6 also has a tendency into this direction. It can therefore be assumed that in Germany and Luxembourg, factor 6 measures also only general attitudes towards immigration and not particular attitudes towards deportation policy as it does in the other countries.

The question remains, why the factor loading structure in Germany and Luxembourg is so different from the rest. In order to find out about this, a change of the abstraction level is indicated, Therefore, analyses on the item level are conducted to find out if a single item could be responsible for the construct bias by functioning differently in these countries.

Comparison of means

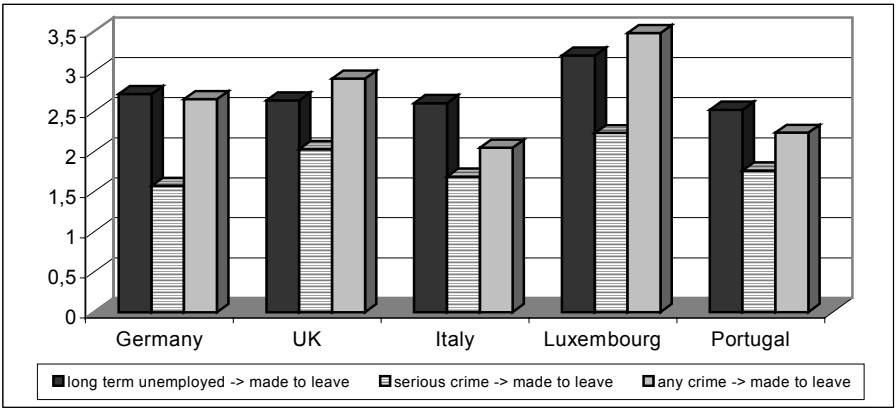
Firstly, a comparison of the means of the three factor 6 items across the five countries can show differences that are not only due to cross-country attitude differences but that are due to a differential functioning of a single item and therefore have to be interpreted as an item bias which then again could have affected the construct equivalence (van de Vijver & Leung, 1997).

In Figure 2 the means of the factor 6 items can be compared. Low means stand for an agreement on the respective item, i.e. immigrants should be made to leave, and high means indicate a tolerant attitude, i.e. immigrants should not be made to leave. The answering scale ranges from 1 to 5. It can be seen that in all countries, the agreement on a deportation of immigrants is highest in case of a serious crime. You can also see that out of all five countries, Luxembourg is the most tolerant one, having the highest means. But it also turns out that in Germany and Luxembourg, the differences between the item "serious crime" and the two others is bigger than in the other three countries[4]. People from Germany and Luxembourg are, in comparison to the UK, Italy and Portugal, much more in favour of making immigrants leave their countries after they have committed a serious crime, than if they only had committed "any crime" or had been "long term unemployed".

3 As too high loadings, I consider factor loadings that are a) higher than .2 and b) deviate with more than .15 from the general loadings.

4 This can be seen more clearly by comparing the relative difference of the mean scores across countries. The ratio of the average of the items "unemployed → made to leave" and "any crime → made to leave" to the item "serious crime → made to leave" is 1.70 in Germany and 1.50 in Luxembourg and therefore higher than in the three other countries where the ratio is always lower than 1.38.

Figure 2 Mean Scores of Factor 6 Items across Countries



Source: ESS round 1, edition 4.1; weighted with dweight

Two reasons for this deviation are possible. Either the true values really differ in Germany and Luxembourg or an item bias has occurred. If the latter is true, then the construct bias of the whole factor also becomes understandable: As the item "serious crime" works differently in Germany and Luxembourg, it also has a lower loading and therefore less impact on factor 6 which then measures more 'general attitudes' than 'detailed attitudes' towards deportation in case of criminal behaviour. Even though it would be easier to find out we have to deal with an item bias if cognitive interviews were available, it is still possible to find explanations for an item bias by taking a look at the society for example.

In Germany, in the course of the discussions about the new immigration law, deportation policy was discussed intensively in the media. Germans also remembered the trouble that the Bavarian government had when they had wanted to send a 14-year old Turkish serial delinquent back to Turkey. Similar reasons might also apply to multicultural Luxembourg that seems to be quite sensitive towards everything concerning safety policy.

Another reason for an item bias of the "serious crime" item and consequently also the construct bias of the whole factor 6 in Germany and Luxembourg might be found by checking the item formulation and translation itself.

Check of linguistic equivalence

The English source questionnaire uses the wording "If people who have come to live here commit a serious crime, they should be made to leave" which can easily be understood as meaning an active deportation by the government.

But in German, the item formulation is "Wenn Zuwanderer, die nach Deutschland kommen, eine schwere Straftat begehen, dann müssen sie das Land wieder verlassen." This is much weaker and only has the meaning of "delinquents have to or must leave the country". Here it remains open how they should leave – if the government has to deport them actively or if the delinquents themselves should be reasonable enough to leave the country. This difference in meaning also applies to the other 2 factor-6-items. So once you realise that the German questionnaire did not ask explicitly for the people's opinion on an active deportation by the government, it becomes understandable that factor 6 has much in common with factor 2 and 7, both of which measure attitudes towards immigration on a more general level.

In the translation for Luxembourg, "Si les gens, qui sont venus vivre ici, commettent une infraction grave, il faut les obliger à repartir.", there does not seem to be any semantic differences to the wording in the source questionnaire, so it seems that the item is linguistically equivalent.

To find out which item bias possibly occurred in Luxembourg and to confirm the hypothesis about the reasons for the construct bias in Germany, some cognitive interviews are needed.

4.4 Structural equivalence of factor 5 – "cultural characteristics"

As Table 1 showed, it is Italy and again Luxembourg that have a low score of factorial agreement on factor 5, which measures attitudes towards cultural characteristics of immigrants. Possible reasons for this construct bias will now be looked at, using the same methods as for the analysis of the construct bias of factor 6.

Comparison of factor loadings

The upper part of Table 3 again shows the factor loadings of the five items associated with factor 5 on that factor. It can be seen that – especially in Italy – there are quite big deviations from the general factor solution. All of the five items show smaller loadings on factor 5. This is also true for two items in Luxembourg, namely the item "Qualification for immigration: be white" and "Better for country if almost everyone share customs and traditions".

Table 3 Factor Loadings of Selected Items on Factor 5

		All countries	ITA	LUX
F5	Better for a country if a variety of different religions	.50	*.28*	.54
	Qualification for immigration: Christian background	.67	*.39*	.65
	Qualification for immigration: be white	.69	*.42*	*.45*
	Better for a country if almost everyone share customs and traditions	-.50	*-.31*	*-.05*
	Qualification for immigration: be wealthy	.56	*.28*	.43
F4	Qualification for immigration: good educational qualifications	.05	*.32*	.12
	Qualification for immigration: speak country's official language	.11	*.36*	.09
	Qualification for immigration: close family living here	.11	*.32*	.22
F3	Immigrant different race/ethnic group majority: your boss	.17	.17	*.37*
	Immigrant different race/ethnic group majority: married close relative	.19	.15	*.46*
F8	Better for a country if almost everyone speaks one common language	-.01	*-.21*	*.21*

Source: ESS round 1, edition 4.1; weighted with dweight; factor loadings higher than .2 and a deviation of more than .15 from the general loadings in italics.

The lower part of Table 3 shows which of the other items show – misleadingly – high loadings on factor 5 in Italy and Luxembourg. In Italy, almost all of the items that in fact should belong to factor 4 show the same high loadings on factor 5 as the factor 5-items. In detail, this applies to the items "Qualification for immigration: good educational qualifications", "Qualification for immigration: speak country's official language", "Qualification for immigration: close family living here" and also the factor 8 item "Better for a country if almost everyone speaks one common language". A hypothesis could then be set up that for Italy no distinction between the qualifications and cultural characteristics can be made. Consequently, this would indicate a construct bias.

This different structure in Italy can also be seen in a confirmatory factor analysis in which covariances of the factors are allowed. Factor 4 and 5 correlate with almost one whereas the covariance in the general solution is rather high with .63, but significantly lower than 1.0. So it seems that in Italy factor 4 and 5 only constitute one single factor in reality. This combination then measures what Italians think immigrants should bring with them and how they should behave.

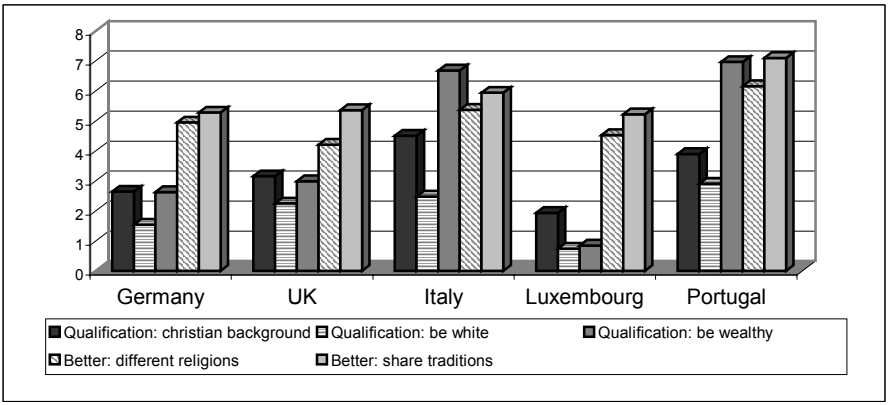
In Luxembourg, other items also have higher loadings on factor 5. These are the three items indicated in the last three lines of Table 3: "Immigrant different race/ethnic group majority: your boss", "Immigrant different race/ethnic group majority: married close relative" and "Better for a country if almost everyone speaks one common language". All of these items measure attitudes in relation to "culture", i.e. how important cultural char-

acteristics and ethnic background for personal contacts are. So, whereas in Italy, factor 4 and 5 form one general factor, it seems that in Luxembourg, factor 5 together with factor 8 and factor 3 constitute one big "culture factor". The hypothesis then is that, as Luxembourg is per se a multilingual country and also a country with a high proportion of highly qualified foreigners, problems of any kind of cultural integration and background are less relevant in social life. They are therefore less separated in people's minds so that no distinction into three distinct factors can be made in Luxembourg.

Comparison of means

A comparison of the mean scores of the factor 5 items in Figure 3 can again show interesting deviations across countries that could have an effect on the construct bias of factor 5 in Italy and Luxembourg. All items were recoded if necessary so that low values indicate a rather tolerant attitude, i.e. that it is not important that immigrants are wealthy, white etc. High means, on the other hand, stand for a less tolerant attitude, i.e. that it is very important that immigrants are wealthy, white etc. In order to facilitate the comparability, all of the items assessed on a 5-point-scale were transformed into a 11-point-scale ranging from 0 to 10.

Figure 3 Mean Scores of Factor 5 Items across Countries



Source: ESS round 1, edition 4.1; weighted with dweight

At first sight, it seems that Luxembourg is quite tolerant whereas Italy and Portugal are not. If the item "share customs and traditions" is compared, no big deviations across countries can be found, hence, no item bias seems to have occurred here. But a compari-

son of the item "be wealthy" shows that Italy and Portugal have very high means, whereas the mean in Luxembourg is very small. Here an item bias is possible. In Luxembourg, the item "be white" again has a very low mean, which could also point to an item bias. So an item bias should be checked for at each of these items. The case of Portugal should also be looked at more closely. Although the factorial agreement of this factor seems to be sufficient, there might be an item bias nonetheless.

Check of linguistic equivalence

As before, a check of the exact item formulations and translations is advisable in order to find out if the hypothesis of an item bias can be supported.

The English source questionnaire uses the following wording: "Please tell me how important you think each of these things should be in deciding whether someone born, brought up and living outside [country] should be able to come and live here. Firstly, how important should it be for them to…be white?" and "…be wealthy?". In Italy, the items were translated as "…avere la pelle bianca" and "avere una buona salute". The French translation for Luxembourg is "…avoir la peau blanche?" and "…être riche?". And finally, in Portugal, the items were translated into "…ser branco?" and "…ser rico?".

In Italian, the translation of "be white" is correct. But obviously there was a mix up: instead of "wealthy", the word "healthy" was translated in the Italian version. Taking this into account, the high mean becomes understandable and it is true that this item is biased and cannot be compared with other countries.

The French and Portuguese translation of "wealthy" into "rich" could also be problematic to some extent. It could be assumed that "wealthy" is likely to be interpreted as having lots of money and assets – which might not be considered as being that important for immigrants who have just arrived and want to make money. "Rich" on the other hand, might be rather considered as just having money, or having just enough money to earn one's living – something which is considered as important for immigrants.

But even if we assume that the French and Portuguese translation of "wealthy" into "rich" is correct, the items seem biased nonetheless. It is no wonder that people in Luxembourg, which is already a rich country and which predominantly receives fairly rich and white immigrants – mainly from within the EU –, say that money alone is not important for people to come and live in Luxembourg. The opposite is true for Portugal. Immigrants coming to Portugal mainly emigrate from former colonies like Cap Verde or Brazil as well as from Africa. They are predominantly poor and coloured immigrants. Taking these different contexts into account, it is risky to draw the conclusion that for people in Lux-

embourg colour does not matter and therefore that they are more tolerant than Portuguese who seem to be quite intolerant. If Luxembourg received poor, coloured immigrants, maybe the attitudes within the population would be different. Now, notwithstanding the fact that the wording is slightly different, there could still be an item bias here because with different social contexts, the item invokes different reactions in the two countries.

5 Conclusion

Two things should have been shown in this paper. First of all, it has to be stated that although a lot has been done in the ESS to avoid all types of bias, it seems impossible to avoid each and every single bias. On an exemplary basis, some of the few construct and item biases in the immigration module of the ESS were shown. It could also be seen that biases on the lower level, i.e. on the item level affect higher levels of equivalence, too. Therefore the great importance of the phase of questionnaire development was hopefully demonstrated.

Secondly, it was also shown, how important it is that every researcher interested in cross-country comparisons, first analyses if the data are equivalent for his purposes before starting substantive analyses.

Now, if any bias is found, there is no need to give up. If severe item biases occurred, the respective item should be excluded for the analysis in the respective country. This has already happened in edition 5 of the ESS with the Italian wealthy-healthy-item. For some countries a different factor structure will have to be assumed due to a construct bias. In the examples of this paper, this would have to be done for Luxembourg for instance, with its big culture factor or in the Italian case, with its combination of factor 4 and 5. If a different factor structure has to be assumed, the factor scores that are often the basis for further analyses should be calculated differently for each country by using confirmatory factor analysis and different models for each country.

In any case, additional data ought to be collected, for example by conducting cognitive interviews or by using split-ballot designs to make sure if or when a bias has to be dealt with and to find out why it has occurred.

References

Berry, J. W., Y. H. Poortinga, M. H. Segall, and P. R. Dasen. 1993. *Cross-cultural psychology. Research and Application.* Cambridge: Cambridge University Press.

Bollen, K. A. 1989. *Structural equations with latent variables.* New York: Wiley.

Braun, M. 2003. "Errors in Comparative Survey Research: An Overview." Pp. 137-142 in *Cross-Cultural Survey Methods,* edited by Harkness, J. A., F. J. R. van de Vijver, and P. Ph. Mohler. Hoboken: Wiley.

Embretson, S. E. 1983. "Construct validity: Construct representation versus nomothetic span." *Psychological Bulletin* 93:179-197.

Johnson, T. P. 1998. "Approaches to equivalence in cross-cultural and cross-national survey-research." Pp. 1-40 in *Cross-Cultural Survey Equivalence*, edited by J. A. Harkness. ZUMA-Nachrichten Spezial 3. Mannheim: ZUMA.

Harman, H. H. 1976. *Modern factor analysis.* Chicago: University of Chicago Press.

Hui, C. H., and H. C. Triandis. 1985. "Measurement in cross-cultural psychology. A review and comparison of strategies." *Journal of Cross-cultural Psychology* 16:131-152.

OECD. 2003. *Trends in International Migration. Annual Report, 2002 Edition.* Rome: OECD.

van de Vijver, F. J. R. 1998. "Towards a theory of equivalence and bias." Pp. 41-65 in *Cross-Cultural Survey Equivalence*, edited by J. A. Harkness. ZUMA-Nachrichten Spezial 3. Mannheim: ZUMA.

van de Vijver, F. J. R. 2003. "Bias and Substantive Analyses." Pp. 207-246 in *Cross-Cultural Survey Methods,* edited by Harkness, J. A., F. J. R. van de Vijver, and P. Ph. Mohler. Hoboken: Wiley.

van de Vijver, F. J. R., and K. Leung. 1997. *Methods and Data Analysis for Cross-Cultural Research.* Thousand Oaks: Sage.

van de Vijver, F. J. R., and Y. H. Poortinga. 1997. "Towards an integrated analysis of bias in cross-cultural assessment." *European Journal of Psychological Assessment* 13:21-29.

van de Vijver, F. J. R., and N. K. Tanzer. 1997. "Bias and Equivalence in Cross-Cultural Assessment: An Overview." *European Review of Applied Psychology* 41:263-279.

THE LEVEL OF EQUIVALENCE IN THE ISSP 1999 AND ITS IMPLICATIONS ON FURTHER ANALYSIS

VLASTA ZUCHA

Using the International Social Survey Programme (ISSP) for cross-national research high comparability of survey data is expected. The level of equivalence of the secondary data and its implications on research design have to be considered. In this paper equivalence of attitudes towards social inequality in the ISSP is tested by means of confirmatory factor analysis for Austria, Czech Republic and Germany. The research focuses on structural equivalence in cross-national research and examines whether the empirical construct is universal for the countries under investigation. The optimal level of equivalence for testing theories and understanding social realities on the empirical basis of the ISSP 1999 will be addressed. This study serves as example for the linkage of research design, level of equivalence and possible further analysis of quantitative data. For most studies in social sciences structural equivalence of variables might be adequate, depending on the purpose of an empirical study using the ISSP 1999 and on its research design.

1 Introduction

The rising interest in cross-cultural and cross-national research has been manifested in establishing various international survey programs. Some of them exist since the 1970s or 1980s, e.g. the Eurobarometer, the World Value Survey and the International Social Survey Programme. Furthermore new programs like the European Social Survey are established.

The ISSP is a large-scale survey program covering various topics over time and different nations. The data offers extensive opportunities for the analysis of different social phenomena. However, some users of the ISSP still do not pay enough attention on comparability of data. Procedures for testing comparability are seldom applied when using cross-national data for secondary analysis, because testing of several forms and levels of equivalence is time- and cost-intensive. In the context of comparability, implications on the research design have to be considered while using survey data for cross-national research.

The aim of the present paper is to show the level of equivalence achieved in the ISSP 1999 for three countries (Austria, Czech Republic and Germany) and to draw consequences for substantial analysis of the data. First, basic issues of cross-cultural research and standardised surveys are introduced. Second, the concept of equivalence in the context of cross-cultural survey research is discussed and an adequate definition of equivalence is chosen. The research question aims at construct equivalence in cross-national research. It examines whether the structure of the attitudinal variables is comparable for the countries under investigation. Relevant aspects of research design are linked to different levels of equivalence. Then the testing of structural equivalence of the ISSP-data is described. Finally, on the basis of this results conclusions on the usage of the ISSP 1999 are drawn. Types of analysis which can be applied on this data are deduced and implications on the research design for studies using this data are drawn.

2 Cross-Cultural Surveys and the Concept of Equivalence

In cross-cultural and cross-national research the comparison of nations and cultures gives opportunity e.g. for describing social phenomena in different groups or for testing theories in different settings. There are various possibilities for research design and several alternatives of how to treat "country" or "culture" in this context. High attention has to be paid to the implications of research design and the chosen type of study. In this part of the paper the purpose of a cross-cultural study will be linked to requirements in terms of comparability.

2.1 Research design and types of studies

Numerous typologies of studies as well as of analysis have been developed and described (Przeworski & Teune, 1982; Rokkan, 1972; Teune, 1990). The aim of the present paper is to investigate comparability of survey questions and to give recommendation for the use of the ISSP 1999. A basic typology of studies seems adequate. Generally it can be distinguished between studies which are designed for discovering similarities and studies primarily designed for finding differences between countries. To reproduce these general purposes and with regard to the basic issues of cross-national comparability in survey research the typology applied by Alwin et al. (1994) is chosen. It is a simplified version of the typology by Kohn (1989). Accordingly, there are two general principles and possibilities for using cross-national surveys – nation as *object* of analysis on the one hand and nation as *context* variable or as *unit of analysis* on the other hand.

Using nation as *object* of analysis the researcher is interested in the countries themselves and differences between countries. In this type of studies often descriptive information on the countries or their institutions is collected. If nation is applied as *context* variable or as

unit of analysis, social institutions and structures of the countries are investigated. Furthermore, relationships between social phenomena and different institutions are examined. Therefore, a classification of countries by different dimensions and potentially influencing analytic variables can be established to investigate the effects of context (Alwin et al., 1994). Obviously, this twofold typology of studies is analytic and in empirical research mixed types of studies are found.

The decision on how "country" is used in cross-national research and therefore which types of comparisons are made, has methodological implications on the demand of different levels of comparability of concepts and measurement. This results in strategies and standards which have to be followed during the whole research process. If data is used for secondary analysis (like the ISSP) it has to be decided for which purpose they are adequate. Before linking type of study and quality of data, the decision on the typology of equivalence in the context of this paper will be described.

2.2 The concept of equivalence

In addition to reliability and validity as basic issues of quality in national studies equivalence is discussed as another requirement in international research. Equivalence is a key concept and core-requirement in comparative and in cross-cultural research.

In literature, equivalence has been conceptualised in different ways. Many forms of equivalence are described and various definitions exist. Johnson (1998) found more than 50 different terms of equivalence and subsumed them into two different categories – *interpretive equivalence* and *procedural equivalence*. The former deals with similarities on the theoretical and interpretative level of concepts and considers if concepts can be meaningfully compared across different cultures. The latter refers to comparability of methods, measurement and administrative procedures.

In the present paper, the focus lies on the comparability of measurement instruments in international surveys and on equivalence of data which can be tested within the limits of secondary analysis. Thus, the focus lies on procedural equivalence according to the typology of Johnson (1998). As procedural equivalence includes all types of equivalence which refer to measurement, it is necessary to specify those forms of equivalence which can be tested in the phase of data analysis in more detail. In the present work a definition of van de Vijver (1998), who distinguishes three forms of equivalence on the level of measurement, is used:

"Equivalence refers to the measurement level characteristics that apply to cross-cultural score comparisons; three types of equivalence are defined: construct (identity of construct

across cultures), measurement unit (identity of measurement unit), and scalar equivalence (identity of measurement unit and scale origin)." (van de Vijver, 1998: 41)

These three types constitute a hierarchical scheme with *construct equivalence* (also known as and related to functional or structural equivalence) as the basic form of comparability and the lowest level of equivalence. It means that similar constructs are measured in each cultural group and describes the identity of constructs across cultures. *Measurement unit equivalence* is the next level of equivalence which is established if the measurement unit of the instrument is identical for each of the cultural groups. The highest level – *scalar or full score equivalence* – is reached only if the measurement scales have the same origin in all compared cultural groups (van de Vijver & Leung, 1997; van de Vijver, 1998; van de Vijver, 2003). This concept derives from psychometric research and therefore poses high requirements on survey data.

2.3 Research design and level of equivalence

While conducting a primary survey different strategies have to be applied during the whole research process to establish equivalence (Alwin et al., 1994; Harkness, 1998; Niedermayer, 1997; Przeworki & Teune, 1982). In different phases, for instance in the phase of translation or pretesing, some forms of equivalence can be tested by means of cognitive tests or statistical methods. In contrast, analysis of secondary data is limited to tests of equivalence in the phase of data analysis. Data is already collected and the researcher involved in secondary data analysis has to rely on documentation of the primary researcher or institute. In this case, equivalence can be tested ex post. Nevertheless the possibilities of testing are reduced in comparison to possibilities of testing comparability and establishing equivalence during a primary survey process.

Generally, it depends on the research design and the type of study, which level of equivalence has to be determined. If the empirical basis of a research project is secondary survey data, the level of equivalence has to be tested. This has to be considered, because there are wide consequences for further analysis of the data. The researcher has to conclude which type of analysis is appropriate, which kind of empirical work can be applied to the data and therefore which type of study is possible. This section will link and summarise the aspects of research design and levels of equivalence discussed so far.

The optimal level or form of equivalence depends on the research design and the purpose of the study. Tests of equivalence are time and cost intensive, therefore the level of comparability should be considered with regard to the type of comparison conducted in the particular study. Researchers interested in testing theories need another level of equivalence of data than researchers comparing social indicators between countries.

Many empirical studies use country as object of analysis. The purpose of these studies is primary descriptive and focuses on country-specific differences. If based on survey data, the type of analysis is variable-oriented and often frequencies and means are compared. This kind of studies is called *level-oriented*. Descriptive statistics or t-tests are the statistical methods applied in the phase of data analysis. In cross-cultural and cross-national research measurement equivalence is necessary if level-oriented studies and analysis are conducted.

On the contrary, if the primary research interest of cross-national studies lies on testing theories and relationships, the analysis is *structure-oriented*. Nation is applied as context variable or as unit of analysis to examine the generality of assumptions and results as well as relationships between social phenomena and institutions. The studies focus on similarities across countries. Analysing survey data in this context various structure-exploring or structure-confirming methods are applied, e.g. factor analysis, latent class analysis, cluster analysis, structural equation modelling and other structure-oriented statistical methods.

An overview of the twofold typology of studies and analysis as well as the linkage of the level of equivalence is given in Table 1. Different levels of comparability are demanded according to different types of studies and analysis.

Table 1 The Purpose of a Study and the Level of Equivalence

Principles of using surveys	Purpose of study	Type of study and of analysis	Level of equivalence
nation as object of analysis	descriptive, focus on differences	level-oriented studies	measurement equivalence
nation as context/ nation as unit	testing generality, testing relationships	structure-oriented studies	structural equivalence

Certainly, comparability in level-oriented studies is harder to establish, because measurement equivalence is a high requirement on the quality of data. Nevertheless, in social sciences the highest level of equivalence is not always needed, if the examination of relationships of social phenomena and the testing of general theories and assumptions is the purpose of a study.

3 Testing Structural Equivalence

Different methods to test and strategies to establish equivalence of measures during the whole research process have been introduced in literature. Furthermore a variety of methods for the examination of different forms and levels of equivalence in the phase of data analysis and within a secondary analysis exist (van de Vijver & Leung, 1997; Johnson, 1998; van Deth, 1998; Berry et al., 2002).

In this paper the focus lies on the basic level of comparability – on structural or construct equivalence. The examination tries to identify the similarity of structures across countries relying on the definition of van de Vijver & Leung (1997) and van de Vijver (2003). It is possible to test this form of equivalence in the phase of data analysis. Statistical methods which rely on the structure of variables and dimensions are appropriate and can be applied. Most frequently exploratory factor analysis or multidimensional scaling are used for this purpose.[1]

The procedure of testing structural equivalence of attitudinal measures by means of *confirmatory factor analysis* is now described in detail. The research question aims at the structural aspects and dimensionality of variables and examines whether the construct underlying the attitudinal variables is universal for the countries under investigation or not. In comparison to exploratory factor analysis and multidimensional scaling it offers more flexibility and alternatives for modelling, but it is more complex, time-intensive and there are higher demands on sample size and data quality (e.g. on normal distribution of variables and on scale type of the variables).

Confirmatory factor analysis in a multiple group setting was first introduced by Jöreskog in 1971. It is a structure-confirming method which can be conducted simultaneously for several populations and it examines whether the hypothesised relationships between variables and factors can be found in the empirical data. It offers the possibility to test hypotheses about the relationship of variables and underlying dimensions simultaneously for two and more groups. Various fit-indices and a chi-square test offer orientation if and to which extent empirical data reflect the theoretical construct.

Different parameters of the empirical model can be set variant or invariant across groups and alternative models can be evaluated by means of fit-indices and the *chi-square differ-*

1 A detailed description of testing structural equivalence by means of exploratory factor analysis and multidimensional scaling can be found in Braun & Scott (1998), Welkenhuysen-Gybels & van de Vijver (2001) and Fontaine (2003).

ence test comparing chi-square statistics of the investigated groups. In detail, factor loadings, correlations between factors and measurement errors can be modelled across groups.

For this reason, confirmatory factor analysis can determine the extent of structural equivalence. Hierarchical hypotheses can be tested to evaluate the variance or invariance of the factor structure across groups. In general, the researcher can choose between two alternatives for testing structural equivalence: the bottom-up or the top-down procedure (van de Vijver, 2003). The first procedure starts with the assumption of lowest structural equivalence between groups, the second assumes the highest possible structural equivalence.

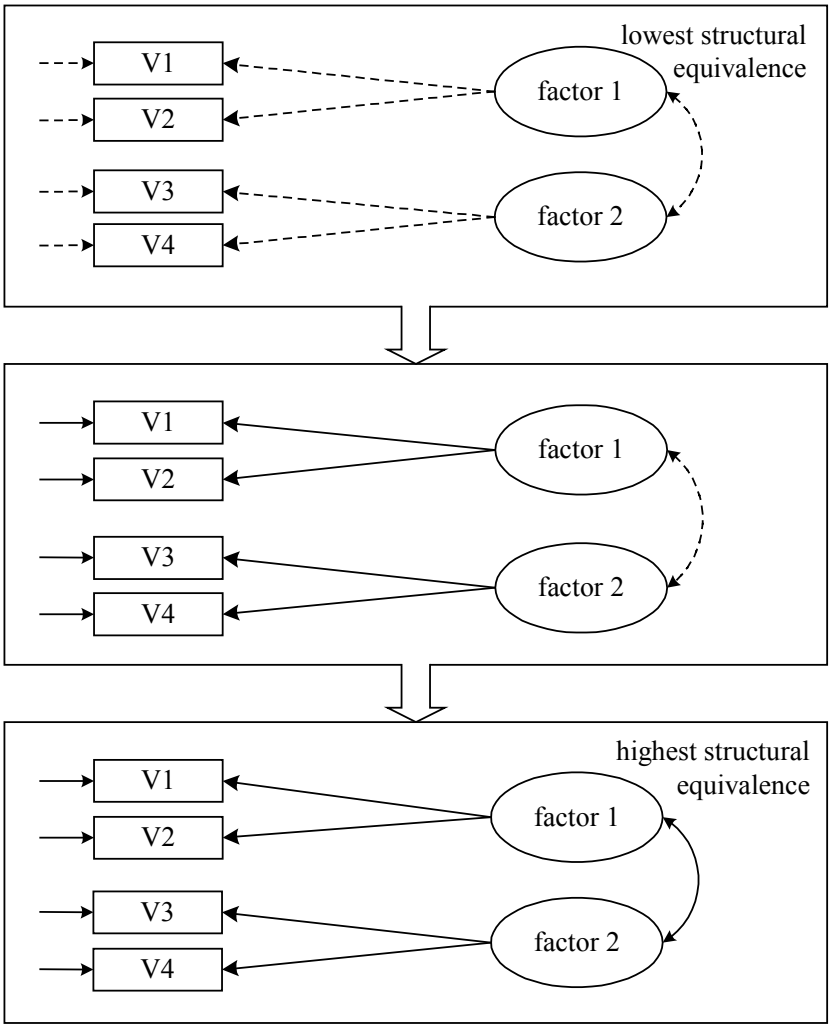
This paper reports on results obtained at the basis of the *bottom-up procedure*. The research questions in this context (indicating the hierarchical hypotheses for the confirmatory factor analysis) can be formulated as follows:

- Is the structure of relationships between items and factors comparable across countries, if factor loadings are not considered?
- If the structure of relationships between items and factors is equal across countries, are the factor loadings and measurement errors comparable across countries?
- If the factor loadings and measurement errors are equal across countries, are the correlations between latent variables comparable across countries?

The first question corresponds to the invariance of relationships and simultaneous variance of factor loadings. It investigates the lowest level of structural equivalence. The second question refers to the invariance of factor loadings and measurement errors. The third hypotheses examines the invariance of factor loadings, measurement errors and correlations between latent variables across all investigated groups or countries. It suggests the highest possible level of structural equivalence. These hypotheses and the process of testing structural equivalence through confirmatory factor analysis is charted in Figure 1 – it shows the process of the bottom-up procedure.

In the present paper structural equivalence of indicators of the ISSP 1999 will be tested. A simplified model of variables measuring attitudes toward social inequality will be introduced. Before testing, basic information on the ISSP and on the investigated items is given.

Figure 1 The Process of Testing Structural Equivalence by Means of Confirmatory Factor Analysis (Bottom-Up Procedure)



Note: The broken line indicates variance across countries in the multiple group analysis, the full line stands for invariance across countries.

4 Testing Structural Equivalence of the ISSP 1999

4.1 The ISSP 1999 and its attitudinal variables towards social inequality

The ISSP is a large-scale survey program conducted since 1983. It covers special topics over time and different nations. The surveys provides a wide range of topics, which are rotating and repeated within a period of several years.

The ISSP 1999 covers the special topic of social inequality and it was conducted in 26 countries. The attitudinal variables cover questions about career advancement by means of family background and networks, social advancement by means of effort, intelligence and corruption, legitimation of inequality, view on earnings and incomes, attitudes towards income inequality, better opportunities through income, social cleavages and conflicts among groups, current and past social position of the respondent, perceptions about and preferences of types of society and social position of the respondents (Harkness et al., 2002). The ISSP provides an opportunity to investigate the perception of social inequality. The possibilities to examine the „objective" and structural level of inequality of a society are very limited, as this is not the primary purpose of the survey.

Testing of equivalence in the present paper is focused on an empirical core model, it does not include all attitudinal variables of the ISSP 1999. Table 2 contains the twelve selected items which are used for the empirical core model to test structural equivalence. The number of the variables in the ISSP-dataset is reproduced in the first column of the table, then the statements and the answering scales are described. The answering scales are five-point scales, asking for agreement/disagreement, justice/injustice or importance/unimportance.

Three countries were selected for the analysis of structural equivalence – Austria, the Czech Republic and Germany. The selection followed the most similar system design introduced by Przeworski & Teune (1982). These countries were selected, because they show a range of similarities, although cultural and historical differences have to be borne in mind:

- Two German-speaking countries contrast one country with a Slavic language. Although Germany and Austria share one language, comparability of measures can not be assumed *a priori*.
- Especially in the context of perception of social inequality the historical and ideological perspective is important. Country-specific differences in the concepts of equality and justice are to be expected, because of the communistic regime forming attitudes and values of the people in the former Czechoslovakia and the former German Democratic Republic for a period of forty years.

Table 2 Selected Items of the ISSP 1999

No.	Statements	Answering Scale
V9	Inequality continues to exist because it benefits the rich and powerful.	agree – disagree
V12	Inequality continues to exist because ordinary people don't join together to get rid of it.	agree – disagree
V34	Differences in income in [COUNTRY] are too large.	agree – disagree
V35	It is the responsibility of the government to reduce the differences in income between people with high incomes and those with low incomes.	agree – disagree
V39	People with higher incomes can buy better health care than people with lower incomes.	just – injust
V40	People with higher incomes can buy better education for their children than people with lower incomes.	just – injust
V50	Important for pay:* The number of years spent in education and training.	important – not important
V51	Important for pay:* Whether the job requires supervising others.	important – not important
V52	Important for pay:* What is needed to support a family.	important – not important
V53	Important for pay:* Whether the person has children to support.	important – not important
V54	Important for pay:* How well he or she does the job.	important – not important
V55	Important for pay:* How hard he or she works at the job.	important – not important

Source: ISSP 1999 Social Inequality III – Final Questionnaire.
* Important for pay: "In deciding how much people ought to earn, how important should each of these things be, in your opinion…?"

It was decided to split the German sample into "East" and "West" (according to "alte und neue Bundesländer") and to treat the three countries as four groups in analysis. Thus the different historical developments of the Eastern and Western part of Germany could be better accounted for. Furthermore, the social and economic heterogeneity within Germany is considered by splitting.

4.2 Testing structural equivalence by means of confirmatory factor analysis

To examine cross-national comparability of attitudes towards social inequality confirmatory factor analysis is conducted simultaneously in all four investigated groups – Austria, Czech Republic, East-Germany and West-Germany. The analysis tries to identify the extent of similarity of structures across countries through setting different parameters variant or invariant.

Data

Confirmatory factor analysis as well as many other statistical procedures are based upon the assumption that variables are normally distributed. Problems of estimation can occur if the distribution of variables departs from multivariate normality. All variables selected from the ISSP dataset for the test of structural equivalence are ordered variables (ordinal level, 5-point scales). The tests for bivariate normality proved, that the normality assumption does not hold for the analysed variables. Ordered and non-normally distributed variables require special handling in structural equation modeling. Therefore, the analysis is based on polychoric correlations and asymptotic covariance matrices. For mathematical reasons, this procedure can be applied only on complete cases and listwise exclusion of missing values is obligatory.

Listwise deletion of missing values reduces the sample size in all countries by 20 percent on the average. This leads to a sample size of 432 respondents in East-Germany and 679 respondents in West-Germany. The Czech sample is reduced to 1.479 and the Austrian sample to 767 complete cases. The analysis of structural equivalence is provided with unweighted data. The parameters were estimated using the WLS-method (Weighted Least Squares). The confirmatory factor analysis was performed through LISREL 8.30, the matrices were produced with PRELIS 2.30.

The basic empirical model

Confirmatory factor analysis is used to test a hypothesised model so the researcher needs prior knowledge and hypotheses about relationships among variables and factors. For testing of structural equivalence of the ISSP an empirical model reproducing the underlying dimensions of attitudes towards social inequality was established. The model was specified due to considerations in literature and exploratory factor analysis.

Two concepts were taken as basic dimensions of the empirical model – *egalitarianism* and *individualism*. The twelve selected variables can be allocated according to this basic distinction of attitudes towards social inequality and (income) distribution. The results of an exploratory factor analysis show that the perception of egalitarianism and individualism is additionally split on two levels – the macro or group level and the micro or individual level. The combination of these dimensions results in four factors:

- Egalitarianism on macro or group level (EGAL_macro)
- Individualism on macro or group level (IND_macro)
- Egalitarianism on micro or individual level (EGAL_micro)
- Individualism on micro or individual level (IND_micro)

The first factor refers to inequality and income differences on societal and group level – in the following analysis the abbreviation EGAL_macro is used for this factor. The second dimension represents better individual opportunities through higher income which are manifested on a societal macro level (IND_macro). The following two dimensions are allocated on the micro or individual level – one referring to egalitarian views on earnings and income (EGAL_micro), the other to individualistic attitudes towards the allocation of income (IND_micro).

The hypothesized model of attitudes towards social inequality in the ISSP 1999 includes four latent variables and twelve manifest variables, but does not reach an acceptable fit in confirmatory factor analysis. Therefore it was slightly modified in such a way that the empirical model now consists of six latent variables and twelve manifest variables. The specification of the modified model is shown in Figure 2 which refers to the manifest variables, latent constructs and their interrelations.

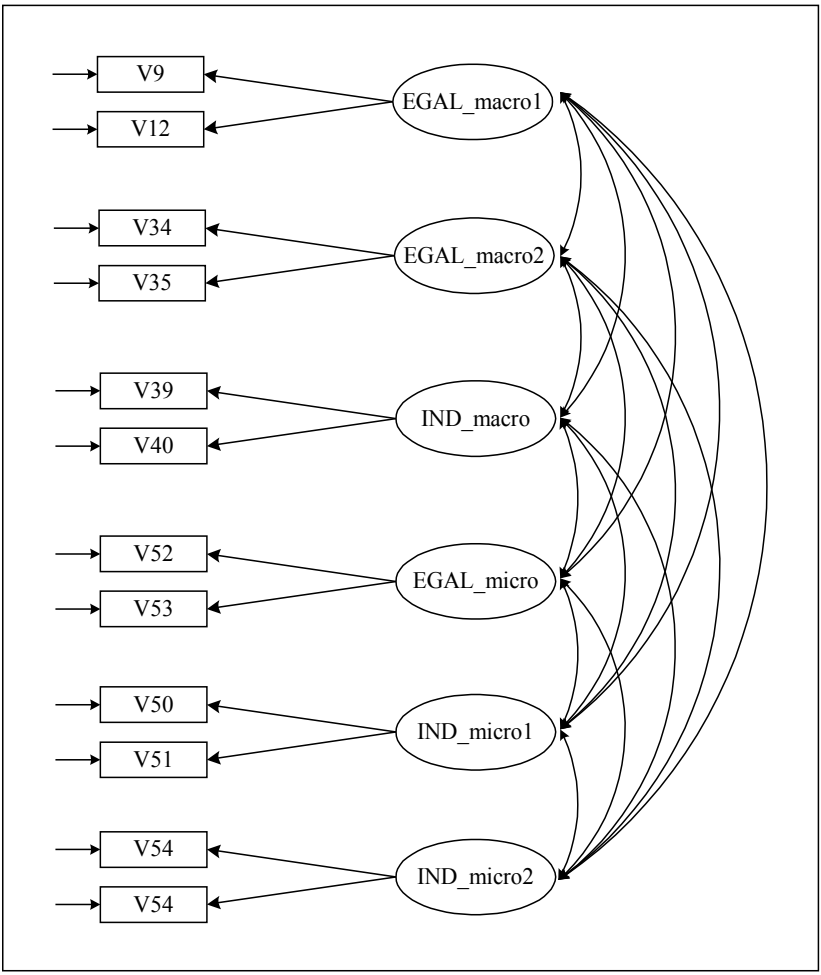
The variables V9 and V12 serve as indicators for egalitarianism on the macro level (EGAL_macro1), as well as the variables V34 and V35 (EGAL_macro2). The factor IND_micro was also split. The items V39 and V40 represent IND_macro and the questions V52 and V53 measure EGAL_micro. All latent variables are assumed to correlate with each other, measurement error variances are not correlated. In structural equation modelling, every latent variable has to be scaled. To determine the scale of the latent variables in the examined model, the latent variables were standardised (no reference variable was used).

The chi-square, degrees of freedom and fit-indices for all investigated groups are given in the table below.

Table 3 Fit of the Basic Empirical Model

country	χ^2	df	p	AGFI	CFI	RMSEA
Austria	66,520	39	,00391	,989	,997	,0304
Czech Republic	112,928	39	,00000	,990	,996	,0358
Germany East	68,868	39	,00222	,985	,995	,0422
Germany West	68,593	39	,00283	,988	,997	,0335

Figure 2 The Basic Conceptual Model for Testing Structural Equivalence By Means Of Confirmatory Factor Analysis



Testing structural equivalence – the results

In this section, confirmatory factor analysis in the multiple group setting will explore whether the same measurement model holds in the four groups (Austria, Czech Republic, East-Germany and West-Germany). Different parameters of the model will be set variant or invariant. In the present paper the *bottom-up procedure* is applied, therefore the test starts with the least restrictive hypothesis. The hierarchical hypotheses are formulated as follows:

- *H-form:* This basic and least restrictive hypothesis tests the common form of the factor model, where the invariance of the form and variance of parameters is assumed. In other words, the number of factors remains the same and the fixed, free and constraint parameters are set in same way across countries. If this common factor structure indicates an acceptable model-fit, the hypothesis holds and the next restrictive hypothesis can be tested.
- *H-load:* In this step the assumption of invariance of factor loadings and measurement error variances across countries is examined – these parameters are set equal in all groups.
- *H-structure:* Further invariance constrains are set on the correlations between the latent variables. The factor loadings, measurement error variances and correlations between factors are equal. If this assumption holds the highest level of structural equivalence is achieved.

The hierarchy of invariance refers to invariance of form and invariance of parameters in consecutive steps. To evaluate the progress the models and hierarchical hypotheses are compared with the *chi-square difference test*. This test compares the less restrictive model with the more restrictive model and demonstrates if the more restrictive hypothesis should be accepted or rejected.

Starting with the assumption of the same form, but variant paths for all countries the chi-square difference test was performed for the factor model of the ISSP 1999 described in Figure 2. The results of the multiple group confirmatory factor analysis of the first model (H-form) is compared to the model corresponding to hypothesis H-load. In the model H-form the structure is equal across all groups (Austria, Czech Republic, East- and West-Germany) and all parameters are variant. In the model H-load factor loadings and measurement error variances are set invariant over all groups. Fit-indices, chi-square, degrees of freedom and the chi-square difference test are summed up in Table 4. The chi-square difference test is significant (χ^2Diff = 129,140; dfDiff = 72) and indicates that the factor loadings and measurement error variances should not be set equal across all groups.

Table 4 Chi-Square Difference Test: Multiple Group Confirmatory Factor Analysis

hypothesis	GFI	CFI	RMSEA	χ^2	df	p[a)]	χ^2Diff	dfDiff	Sign.[b)]
H-form	,995	,996	,0351	316,906	156	,00000	-	-	-
H-load	,993	,994	,0338	446,047	228	,00000	129,140	72	s.

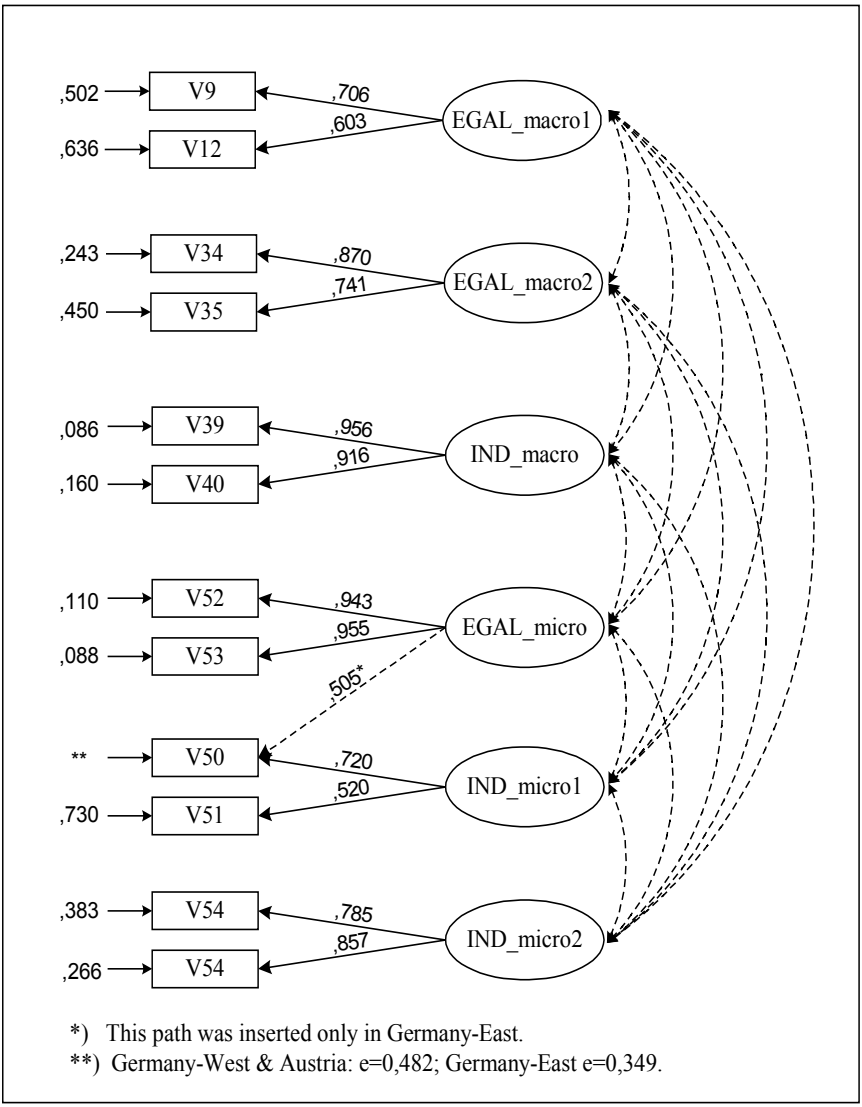
[a)] The p-value indicates, if data significantly differ from the model.
[b)] This column shows if a model differs significantly from the less restrictive model ("s." means "significant", "n.s." means "not significant").

On the one hand, the model of H-form shows an acceptable model-fit, on the other hand the model corresponding to H-load is significantly worse than the first model. For this reason testing has to go into more detail. Consequently, the extent of structural equivalence "lies in between" and the testing can be proceeded for different groups of countries and/or different dimensions.

Various models which lie between the previously tested H-form and H-load were calculated and compared. Different post hoc model modifications were performed by setting systematically specific paths variant and invariant. Finally, an acceptable model with a high level of equivalence was found including the groups Austria, East- and West-Germany. The form of the final model remains the same as of the previous model, only in East-Germany one supplementary path was set which is not replicated in the other two groups. The basis hypothesis is now referred to as H-form2. The more restrictive hypothesis is H-load2 and the most restrictive hypothesis indicating highest level of equivalence is H-structure2.

The parameter estimates (factor loadings and measurement error variances) of the final model (H-load2) are reported in Figure 3, which shows the extent of structural equivalence of the tested model of the ISSP 1999. Factor loadings and measurement error variances are equal in Austria, East-Germany and West-Germany. Only in East-Germany a supplementary path is set between the variable V50 and the factor EGAL_micro.

Figure 3 Structural Equivalence between Germany-West, Austria and Germany-East (Hypothesis "H-load2")



*) This path was inserted only in Germany-East.

**) Germany-West & Austria: e=0,482; Germany-East e=0,349.

This paper reports the results of the final model which shows the highest possible number of invariant parameters in the three groups Austria, East-Germany and West-Germany. Fit-indices, chi-square, degrees of freedom and the chi-square difference test of the models corresponding to the hypotheses H-form2 and H-load2 are reported in Table 5. The test indicates that the chi-square difference is not significant and the model H-load2 is an acceptable alternative.

Table 5 Chi-Square Difference Test: Multiple Group Confirmatory Factor Analysis of the Modified Model (V50 – EGAL_micro), Germany-West, Austria and Germany-East

hypothesis	GFI	CFI	RMSEA	χ^2	df	p	χ^2Diff	dfDiff	Sign.
H-form2	,992	,997	,0326	192,968	116	,00001	-	-	-
H-load2	,989	,997	,0275	240,194	163	,00008	47,226	47	n.s.
H-structure2	,983	,994	,0360	349,078	193	,00000	108,884	30	s.

Level of structural equivalence

The extent of structural equivalence varies depending on which countries or groups are compared. Factor structures of Austria and West-Germany are highly equivalent. Only a few parameters – correlations between some latent variables – have to stay invariant. As well, high structural equivalence is found between Austria, East- and West-Germany. Although, in East-Germany one factor of the model (EGAL_micro) seems under-identified.

Comparing four groups tested in this paper, highest structural equivalence can not be assumed. The four groups (Austria, Czech Republic, East- and West-Germany) show structural equivalence, although the lowest level. Only the form of the factor model is equal and therefore lowest structural equivalence is assumed. The data should be analysed and compared carefully across all four groups (see below). Furthermore, the Czech factor model seems to be under-identified.

5 Implications on Usage of the ISSP 1999

A secondary analysis of the ISSP-data from 1999 was conducted, in which equivalence of attitudes towards social inequality was tested for four groups – for Austria, the Czech Republic, East- and West-Germany. After choosing the framework of structural equivalence confirmatory factor analysis was applied to test cross-cultural comparability of measures and data. Research focused on structural equivalence in cross-national research and examined whether the construct is universal for the countries under investigation. Lowest structural equivalence was found between all countries, highest structural equivalence of the examined ISSP-data can be assumed between Austria, East- and West-Germany.

Before drawing conclusions, some limitations of the presented study should be mentioned. Construct equivalence was tested only in the framework of secondary analysis, therefore it has to focus on a relative small number of variables and rely on methods which can be applied after data collection. Then, the evaluation was conducted only by means of statistical methods – other forms of evaluation could be applied in addition, e.g. cognitive methods.

Conclusions and implications on the usage of the ISSP 1999 can be drawn if the three countries or four groups are involved in a comparative, cross-national study design. First, the analysis of construct equivalence suggests structure-oriented analysis to be appropriate. If level-oriented analysis is applied on this data descriptive measures of the attitudinal variables should not be compared between all countries. Especially the Czech data is limited to the lowest level of construct equivalence and therefore, measurement equivalence cannot be assumed. Only highest level of measurement equivalence permits to compare level-oriented measures across groups. Possible types of analysis which can be applied on the ISSP-data depend on which countries are compared.

Second, the limitations with regard to type of analysis imply conclusions on the type of study and on research design. Depending on the examined countries the researcher can use nation as context variable as well as nation as unit. This kind of study can be conducted with all countries analysed in this paper – Austria, Czech Republic and Germany. Whereas, if a research design is chosen where nation is used as object of analysis, attention has to be paid to the Czech Republic. Measurement equivalence is not established in the ISSP-data for this country, therefore cross-national comparison of frequencies, means and other descriptive, level-oriented measures should be avoided or presented carefully.

To sum up, the testing of attitudinal questions of the ISSP 1999 detected (at least the lowest level of) structural equivalence in three countries. Survey data shows the optimal level of equivalence for testing theories and understanding complex social realities. Depending on the purpose of an empirical study and research design, for most studies in social sciences structural equivalence might be adequate and sufficient.

References

Alwin, D. F., M. Braun, J. A. Harkness, and J. Scott. 1994. "Measurement in Multi-National Surveys." Pp. 26-39 in *Trends and Perspectives in Empirical Social Research,* edited by Borg, I., and P. Ph. Mohler. Berlin, New York: Walter de Gruyter.

Berry, J. W., Y. H. Poortinga, M. H. Segall, and P. R. Dasen. 2002. *Cross-Cultural Psychology. Research and Applications.* Cambridge, New York, Port Melbourne, Madrid, Cape Town: Cambridge University Press.

Braun, M., and J. Scott. 1998. "Multidimensional Scaling and Equivalence: Is having a job the same as working?" Pp. 129-144 in *Cross-Cultural Survey Equivalence,* edited by Harkness, J. A. ZUMA-Nachrichten Spezial, Nr. 3. Mannheim: ZUMA.

Fontaine, J. 2003. "Multidimensional Scaling." Pp. 235-246 in *Cross-Cultural Survey Methods,* edited by Harkness, J. A., F. J. R. van de Vijver, and P. Ph. Mohler. Hoboken, New York: Wiley Interscience.

Harkness, J. A. (Ed.). 1998. *Cross-Cultural Survey Equivalence.* ZUMA-Nachrichten Spezial, Nr. 3. Mannheim: ZUMA.

Harkness, J. A., E. Scholz, and S. Stadler. 2002. *ISSP 1999 Germany. Social Inequality III.* ZUMA Report on the German Study. ZUMA Methodenbericht.

ISSP 1999. *Social Inequality III, Final Questionnaire.* ZA No. 3430, secretariat version.

Johnson, T. P. 1998. "Approaches to Equivalence in Cross-Cultural and Cross-National Survey Research." Pp. 1-40 in *Cross-Cultural Survey Equivalence,* edited by Harkness, J. A. ZUMA-Nachrichten Spezial, Nr. 3. Mannheim: ZUMA.

Jöreskog, K. G. 1971. „Simultaneous factor analysis in several populations." *Psychometrika* 36:409-426.

Kohn, M. L. 1989. „Introduction." Pp. 17-31 in *Cross-national research in sociology,* edited by Kohn, M. L. Newbury Park, London, New Delhi: SAGE.

Niedermayer, O. 1997. "Vergleichende Umfrageforschung: Probleme und Perspektiven." Pp. 89-102 in *Vergleichende Politikwissenschaft. Ein einführendes Studienhandbuch,* edited by Berg-Schlosser, D., and F. Müller-Rommel. Opladen: Leske + Budrich.

Przeworski, A., and H. Teune. 1982. *The logic of comparative social inquiry.* Malabar, Florida: Krieger.

Rokkan, S. 1972. *Vergleichende Sozialwissenschaften. Die Entwicklung der inter-kulturellen, inter-gesellschaftlichen und inter-nationalen Forschung. Hauptströmungen der sozialwissenschaftlichen Forschung.* Frankfurt/Main, Berlin, Wien: Ullstein.

Teune, H. 1990. "Comparing Countries: Lessons Learned." Pp. 38-62 in *Comparative methodology: Theory and practice in international social research,* edited by Øyen, E. London: Sage.

van de Vijver, F. J. R. 2003. "Bias and Substantive Analysis." Pp. 207-233 in *Cross-Cultural Survey Methods,* edited by Harkness, J. A., F. J. R. van de Vijver, and P. Ph. Mohler. Hoboken, New York: Wiley Interscience.

van de Vijver, F. J. R. 1998. "Towards a Theory of Bias and Equivalence." Pp. 41-65 in *Cross-Cultural Survey Equivalence,* edited by Harkness, J. A. ZUMA-Nachrichten Spezial, Nr. 3. Mannheim: ZUMA.

van de Vijver, F. J. R., and K. Leung. 1997. *Methods and Data Analysis for Cross-Cultural Research.* Thousand Oaks, London, New Delhi: Sage.

van Deth, J. W. 1998. "Equivalence in comparative political research." Pp. 1-19 in *Comparative Politics. The problem of equivalence,* edited by van Deth, J. W. London, New York: Routledge.

Welkenhuysen-Gybels, J., and F. J. R. van de Vijver. 2001. *Methods for the Evaluation of Construct Equivalence in Studies Involving Many Groups.* http://www.kuleuven.ac.be/dvz/softwaredata/webversion.pdf (2003-08-27).

PART III: HARMONISING SOCIO-DEMOGRAPHIC INFORMATION IN DIFFERENT TYPES OF SURVEYS

HARMONISING SOCIO-DEMOGRAPHIC INFORMATION IN HOUSEHOLD SURVEYS OF OFFICIAL STATISTICS

Experiences from the Federal Statistical Office Germany

THOMAS KÖRNER & IRIS MEYER

In order to provide the European Union with comparable statistical information, the European Statistical System (ESS) has developed different strategies of cross-national harmonisation. These strategies range from a complete (input) harmonisation of concepts and survey methodology to output harmonisation approaches mainly taking place on the level of the aggregated data. Such strategies challenge statistical agencies in different ways and they have specific strengths and weaknesses. The paper focuses on the harmonisation of socio-demographic information in household surveys of official statistics. We discuss the approaches used in the new Community Statistics on Income and Living Conditions (EU-SILC).

1 Introduction

The availability of harmonised data on income and living conditions in the member states of the European Union is on top of the political agenda. Decisions in the field of social policy in the 25 member states are highly dependent on an accurate and comparable data basis. The Lisbon European Council of March 2000, for example, defined the eradication of poverty as one of the highest political priorities. The political strategy chosen necessitates reliable statistical information: According to the heads of government of the member states the national policies for combating poverty and social exclusion shall be co-ordinated on the basis of the so called open method of co-ordination "combining common objectives, national action plans, common indicators with the aim of promoting more ambitious and effective policy strategies for social inclusion" (Commission of the European Communities, 2003: 4). Official statistics constitute the basis for the definition of these objectives and the monitoring reports.

With the Community Statistics on Income and Living Conditions (EU-SILC) – the successor of the European Community Household Panel (ECHP) – a new EU-wide data source has been created which feeds a whole system of reports and indicators. These reports include the annual Report to the Spring Council, the annual Social Situation Report, the periodic Joint Social Inclusion and Social Protection Report, and others.

While the political importance of harmonised data on income and living conditions is evident, the approaches towards the production of harmonised data are complex. A precise measurement not only of the survey variables but also of the socio-economic background variables is necessary. Harmonised measurement depends on various factors, such as a common definition, taking into account cultural differences affecting the meaning of variables, as well as an operationalisation of the variables respecting the specific national contexts. Harmonisation thus requires a careful consideration of both survey concepts and survey methods.

This paper focuses on problems of the harmonisation of socio-economic variables in household surveys of official statistics. We first summarise the main strategies of harmonisation in the European Statistical System (ESS). In a second step, we outline the harmonisation efforts in the field of surveys on income and living conditions, where a change from input harmonisation to ex-ante output harmonisation can be stated. Finally, we take a closer look at the harmonisation of socio-economic variables in the context of the EU-SILC which is carried out in Germany beginning in 2005. The harmonisation strategy for selected socio-demographic variables is analysed in detail.

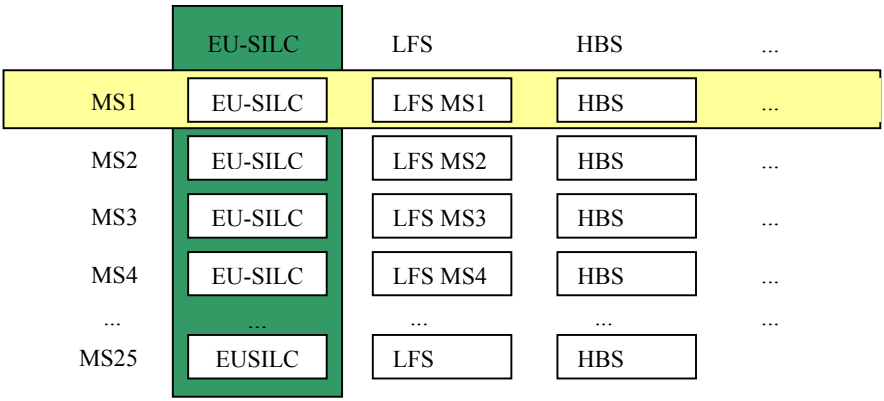
2 Harmonisation Strategies in the European Statistical System

Providing harmonised European statistics is an essential function of the European Statistical System. The notion of harmonisation appears in almost all basic legal documents referring to official statistics. The Quality Declaration of the European Statistical System heralds improving "a programme of harmonised European statistics that constitutes an essential basis for democratic processes and progress in society" (Eurostat, 2002: 21) as the central "vision" of the ESS. However, any approach towards cross-national harmonisation is limited by two features of the ESS institutional environment: The subsidiarity principle and the national statistical systems. Both bring about specific challenges and opportunities.

Firstly, subsidiarity is a key principle of the EU administrative structure. The way European official statistics are produced is heavily relying on the existing structures in the national statistical systems. They encompass far more than only technical or purely administrative features. In the long tradition of the national statistical systems fundamental methodological traditions have been formed. In order to harmonise statistics at EU level successfully, these experiences and traditions must be taken into consideration.

Secondly, official statistics constitute more than a bunch of isolated surveys. In each member state, there are statistical systems which enable users to combine statistical information originating from different surveys. Long before harmonisation became an issue at the European level, these national information systems have taken shape. In order to get used properly, statistics have to fit into this environment. As Mejer (2003: 83) notes, the poor integration of the ECHP in the existing national statistical systems was one of its major drawbacks, thus creating a "potential for fragmented figures showing differing results". Hence, harmonisation in official statistics has at least two dimensions which could be referred to as "cross-national harmonisation" and as "cross-survey harmonisation". Successful harmonisation in one of these dimensions often counter acts harmonisation efforts in the other one. Figure 1 illustrates the problem of both types of harmonisation or selected household surveys. The horizontal bar represents the problem of cross-survey harmonisation in member state 1.

Figure 1 Cross-National and Cross-Survey Harmonisation



MS: Member state
EU-SILC: Community Survey on Income and Living Conditions
LFS: Labour Force Survey
HBS: Household Budget Survey

Cross-survey harmonisation in the national statistical system of MS 1

Cross-national harmonisation for the case of the EU-SILC

The vertical bar represents the problem of cross-national harmonisation in the case of the EU-SILC.

The current level of harmonisation of official statistics on the European level has been achieved by adopting a variety of different harmonisation strategies, tailored to each individual case. Evidently, these harmonisation strategies will not lead to a complete harmonisation of the entire set of concepts and methods used.[1] Scholars like Günther (2004) or Grais (1999) distinguish input harmonisation, ex-ante output harmonisation and ex-post output harmonisation as strategies towards a harmonisation of survey data in the European Statistical System (Figure 2):

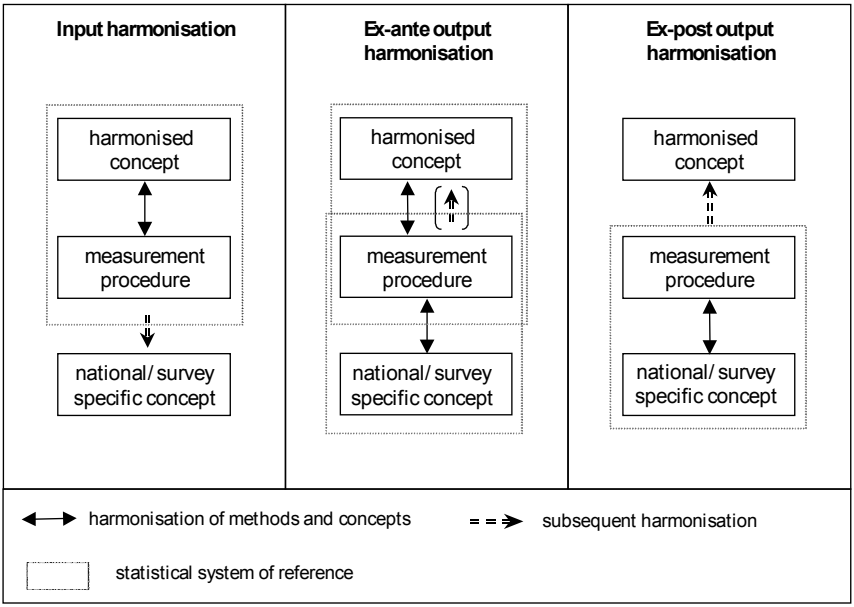
For *input harmonisation* all methods, concepts and procedures relevant for the production of statistical results are standardised from scratch. Therefore, input harmonisation is necessarily applied ex-ante and does in principle not take into account statistical systems already existing in member states. For this reason, substantial changes of the methods, definitions, and processes used in member states become necessary which in turn often lead to increased cost and could hamper the usability of the data on the national level. It is not surprising that there are only few examples for the use of this strategy, such as the European Community Household Panel (ECHP) and the Community Innovation Survey (CIS).

In contrast, with *output harmonisation* standardisation is limited to the statistical results. While the definitions of the statistical results are binding, the National Statistical Offices are (within certain limitations) free to choose the methodology to be applied. One can either implement a new survey in order to meet the requirements of the statistical results; we refer to this strategy as ex-ante output harmonisation. It is used for the EU-SILC. Or one can use pre-existing national data sources and convert the results into a harmonised format. This strategy is referred to as ex-post output harmonisation. Ex-post output harmonisation is supposed to cause lower costs than any other strategy of harmonisation, but its applicability has strict limitations and preconditions fairly difficult to fulfil (Minkel, 2004).

It has to be noted that these harmonisation strategies have to be regarded as ideal types which will not be purely found in the empirical reality of surveys. Even the ECHP has not been fully harmonised as some degrees of freedom were left to the National Statistical Institutes, e.g. regarding specific aspects of the sampling technique. For the EU-SILC, here classified as ex-ante output harmonised, quite a number of common rules for the methodologies to be used have been put in place. Consequently, the EU-SILC could still be regarded as a partly input harmonised survey (Mejer, 2003: 72).

1 One could argue that in different cultural contexts different methodological approaches might, however, be considered functionally equivalent. The same survey mode, e.g. mail surveys, might lead to very different mode effects in different cultural contexts. Such considerations suggest that problems of harmonisation have to be reconsidered in a more differentiated manner.

Figure 2 Harmonisation Strategies (adapted from Günther, 2004: 9)



The harmonisation of socio-economic variables – which is a special case of the harmonisation of statistical surveys in general – can use similar strategies (Hoffmeyer-Zlotnik & Wolf, 2004). In this case, input harmonisation denotes that the harmonised concept is included already in the measurement instrument (e.g. the questionnaire). In the case of output harmonisation, measurement is carried out with reference to a non-harmonised (national or regional) concept, and the harmonisation will take place at a later stage by transforming the non-harmonised results. One would refer to such a strategy as "ex-ante" if the harmonisation is already planned in the design stage of the survey and as "ex-post" in the case of different surveys on the same topic which have been designed independently from one another. These strategies have to be applied complementarily within a given survey, as for each variable an appropriate approach must to be tailored. Some variables (like education) cannot reasonably be measured in an input harmonised ways whereas for others the detour via the measurement of national concepts can lead to important "losses". Section 3 will illustrate this process with reference to selected socio-demographic variables in EU-SILC.

3 Harmonising Socio-Demographic Variables in Surveys on Income and Living Conditions

3.1 From ECHP to EU-SILC

Initially, the ECHP was carried out as an input harmonised survey. The so called "blueprint method" using uniform questionnaires, detailed definitions as well as rules and procedures was applied (Günther, 2004: 17). German official statistics conducted the ECHP from first to the third panel wave (1994 to 1996) in this input harmonised way. From 1997 to 2001, together with Luxembourg and the United Kingdom, Germany changed to ex-post output harmonisation. In this period, the data have been converted from the German Socio-Economic Panel (SOEP) into the format of the ECHP.[2]

In 1999, the ESS decided to change from input to output harmonisation and to replace the ECHP by the EU-SILC. The reasons for this decision were the need for updating the content according to the new political demands and requests for operational improvements, mainly the timeliness of the data (Mejer, 2003: 72). Thus the National Statistical Institutes were given more flexibility in the choice of survey methods. The use of existing national data sources was explicitly encouraged in order to achieve a better integration into the national statistical systems and to make sure that the "best national source" for information on income and social exclusion is used (Eurostat, 2001: 7).

Especially the Nordic countries use register information on income and further selected socio-demographic information. Most of those member states in which register information is not available, decided to implement a new survey for the EU-SILC. This might be interpreted as an indicator for the difficulties that arise when trying to integrate a new and extensive survey like the EU-SILC into an existing national survey (Meyer, 2004).

3.2 Harmonising socio-demographic variables in EU-SILC

While the ECHP was carried out on the basis of a "gentleman's agreement", in the case of the EU-SILC, the harmonisation of socio-demographic variables is prescribed by a number of EU regulations. The regulations provide, e.g. common definitions, the use of standard classifications for all National Statistical Institutes as well as the entire set of target variables and items. They incorporate nearly all units, variables, and classifications which have been defined in the Eurostat recommendations on the harmonisation of "core variables" (Eurostat, 2000).

2 In this contribution we will not focus on the problems of ex-post output harmonisation. The implications of this strategy have been investigated in detail in the CHINTEX project (The Change from Input Harmonisation to Ex-post Harmonisation in National Samples of the European Community Household Panel – Implications on Data Quality; see Ehling & Rendtel et al., 2004).

Given this solid legal framework, the cross-national harmonisation of socio-demographic and other variables might be considered close to an ideal solution as there is a legal obligation to adopt harmonised concepts throughout the entire survey. However, the situation is more complex. As Hoffmeyer-Zlotnik & Wolf (2004: 6) point out, it is only a first step to agree on common definitions. In addition further points, like the cultural and national background of the variables, the operationalisation of the common definition and the wording and design of the measurement instrument (are the concepts easily understood by all respondents?) have to be taken into account. Thus, successful harmonisation has to be prepared carefully already during the survey planning and questionnaire design phases.

The common definitions prescribed by regulations in some cases already have some bearing on the operationalisation (e.g. the household definition), resulting in input harmonised approaches. In other cases, only the target variable is defined, i.e. the breakdown according to which data on a given variable has to be delivered. In these cases the operationalisation is largely left to the National Statistical Institute: Can the target variable be measured as worded in the list of target variables (input harmonisation), or do we have to find a national operationalisation (output harmonisation)? Of course, the national operationalisations of the common definitions have to be functionally equivalent.

In the following we will present a number of selected examples of how the harmonisation of socio-demographic variables has been dealt with in the case of the implementation of the EU-SILIC in Germany.

Private household and relationship of household members

Private households are at the main population unit of the EU-SILC. The concept of household is defined in the regulations in detail, i.e. that input harmonisation has to be applied at least to some extent. The definition of a private household used for the EU-SILC combines the aspects of (a) living together and (b) sharing expenditures. This definition is in line with the one recommended by Eurostat and applied by the member states in the Household Budget Surveys (Eurostat, 2003a: 15). The regulation further specifies this quite broad and general definition in order to have unequivocal criteria of which individuals belong to a household: In addition to the general definition, household members are identified by a full set of conditions for inclusion, for instance

- having no private address elsewhere,
- having spent most of the daily night-rest in the household over the past six months, and
- intending to stay in the household for a period of at least six months.

For specific categories of persons further criteria apply. For example are considered as household members "persons temporarily absent [e.g. in hospital] but having household ties", i.e. having financial ties and being absent for less than six months. However, if people are temporarily absent in private accommodation, they "may be included as household members irrespective of the length of absence, provided they are not considered members of another private household" (Commission Regulation (EC) 1980/2003).

Such a complex definition might be helpful to make sure that the risk of double-counting or under-coverage in the sampling frame are minimised. However, the definition is rather difficult to operationalise, at least if using self-administered questionnaires (as practiced in the German implementation of the EU-SILC). It is hardly possible to inform respondents on all the conditions for being regarded as a household member so that there is a risk that respondents simply rely on the definition on household membership they have in mind from their everyday knowledge. Response will be given (at least in part) according to the culturally shaped conception of private household. In order to solve this problem in the case of Germany, we used the general definition (residence and sharing expenses) and simplified its further specification for the questionnaire by illustrating it by different examples.

Concerning the household structure, in the regulation, there was a change from input to output harmonisation. In the ECHP, household structure was measured by the help of a relationship matrix in which respondents had to name the relationship of every single household member with every other member of the household. In the EU-SILC, this matrix is no longer required; the target variables can also be obtained with an alternative operationalisation. One of the reasons might have been that the relationship matrix can be difficult to respond at least in the case of large households or in self-administered questionnaires. Our experience from the EU-SILC pilot study as well as the pretests for the EU-SILC questionnaires have shown that completing a relationship matrix is a complex and cognitively demanding task which is prone to response errors, item nonresponse, or even a break-off of the interview.[3]

For these reasons the Federal Statistical Office Germany decided to no longer to use the complete matrix of all household members in the EU-SILC. Alternatively, we ask for the relations of a reference person only to each of the household members and add few complementary questions (similar to the concept used in the Labour Force Survey). The complete relationship matrix, in the sense of the ECHP, can then be obtained by calculating the complementary relationships based on the relationships filled in for the reference

3 More generally, recent research has again shown that particular groups of respondents have difficulties to fill in matrices or tables in survey questionnaires (see, e.g. Timm, 2004).

person. The change towards output harmonisation clearly gave the National Statistical Institutes greater flexibility to adopt their "best national practice". Furthermore, given the German experience, a complete household relationship matrix in all different national contexts and traditions had probably not helped to achieve a higher degree of harmonisation. The harmonised method would have been prone to various kinds of errors, including not only measurement errors, but also processing errors (due to the fact that the analysis of the data from the matrix is a time consuming and complex task).

Occupation and employment

For a harmonised measurement of occupations in cross-national surveys, the International Standard Classification of Occupations of 1988 (ISCO-88) is a standardised and widely accepted instrument. Not surprisingly, also the EU-SILC relies on ISCO-88 (on the two-digit level) in order to obtain background information on the current main job. As the ISCO-88 is a highly complex instrument, it is still not considered possible to categorise occupations according to ISCO-88 during the data collection phase (Hoffmeyer-Zlotnik, Hess & Geis, 2004). In principle, input harmonisation could be adopted: Respondents could be asked to provide information on their current job in a set of open questions. Subsequently, the information has to be recoded according to the international classification.

However, in the German case there is a further complication. Besides the ISCO-88, there is also a national standard classification on occupation, the "Klassifizierung der Berufe 1992 (KldB)", published by the Federal Statistical Office. The KldB is still a necessary information resource within the German statistical system. Due to the differing classification principles of the ISCO-88 and the KldB, a conversion from ISCO-88 to KldB is not possible, whereas within certain limitations, KldB information can be converted into the ISCO-88 format.[4] Unfortunately, also the conversion from KldB to ISCO-88 leads to some loss of information which could only be avoided by adopting a strategy of input harmonisation (i.e. phrasing the respective questions according to the requirements of ISCO-88 and coding directly from this information). However such an option would heavily restrict the use of the results within the national statistical system. As different

4 ISCO-88 categorises occupations on the basis of the level and similarity of the skills required. It uses the two dimensions of (in a first step) "skill level" (range and complexity of a task) and (subsequently) the "skill specialisation" (reflecting the type of knowledge applied, tools and equipment used, materials worked on, or with, and the nature of the goods and services produced). ISCO-88 distinguishes nine major groups, which are further subdivided into 390 unit groups on the four-digit level (ILO 1990: 2). The KldB, in contrast, distinguishes occupations according to the occupational title, starting from six "occupational areas" (like "occupations in agriculture, forestry and gardening") all the way down to 29527 concrete occupation titles. On the three-digit level, 369 orders of occupations are distinguished (Statistisches Bundesamt, 1992).

survey questions are required also parallel coding according to both classification systems is not an option. Finally, it should be mentioned that coding according to ISCO-88 is by far more resource consuming.

As for the EU-SILC ISCO-88 codes are required on the two-digit level only the loss of information during the recoding can be restricted to an acceptable level. Thus, a conversion table can be used, which has already been applied in the German Labour Force Survey since 1996 (Emmerling & Riede, 1997; Schwarz, 2001).

This case shows that in official statistics the requirements of international and national users have to be reconciled. There is not a major problem as long as national and international classification systems are identical, at least on a broad level. For example, for the coding of economic activities the Statistical Classification of Economic Activities in the European Community (NACE rev. 1) is identical with the national classification "Wirtschaftszweige 2003" (WZ 2003) on the two digit level. In contrast, the case of the measurement of occupations makes clear that in case of conflicts between national and international classifications systems solutions have to be found which take into consideration the requirements of both national and international users.

Education

As for the occupation, also for education, an international standard classification is available and adopted in EU-SILC – the International Standard classification of Education (ISCED 1997). As the ISCED categories are quite remote from the everyday life of respondents, output harmonisation is the only feasible harmonisation strategy: Respondents are asked to tick the category which applies to them out of a list of national educational levels. In Germany, this information can subsequently be recoded according to ISCED 97 without major problems as a standard conversion table exists.

However, despite the fact that for education the national and international classification systems are convertible, some questions arise with respect to the harmonisation. The national education systems greatly differ across nations, which is very difficult to grasp by an international classification. In contrast to the situation with ISCO-88, for ISCED 97 this does not conflict with a measurement which is also instructive for national purposes. However, ISCED fails to represent cases of horizontally differentiated education programmes. For example, vocational training in Germany cannot be represented appropriately side by side with the general educational levels. Nevertheless such a request would possibly overburden an international classification system.

Income

The main purpose of the EU-SILC is to become the new reference source for EU comparisons on poverty, living conditions, and income distribution. As basis for the so called "open method of co-ordination" the EU-SILC is also one of the major inputs for decisions on social policy within the member states of the EU. This aim necessitates first of all a highly precise and extensive measurement of the income. The major income information required for the EU-SILC includes:

- the total gross household income,
- the total disposable household income, and
- the total disposable household income before social transfers including/excluding old-age and survivor's benefits.

This income information will be compiled on the basis of a broad variety of income components (gross as well as net) on household and personal level. Income reference period is the previous year. The retrospective measurement is necessary as some income components are only available after the reference year and only on a yearly basis (like tax repayment). According to the regulation, a multitude of income components have to be provided in equally high precision: income from self-employment, from employment, from social transfers and from private pension plans. Additionally, income from capital as well as near-cash income have to be reported. The German implementation of EU-SILC comprises 89 types of income information which are required in order to provide the 23 target variables laid down in the regulations.

This brief overview makes clear that for the EU-SILC, income is far more than a socio-demographic background variable. It is the one of the main focus areas of the survey. As a simple background variable, one would not have to measure income in such a differentiated way. For this reason, the problem of a harmonised measurement of income differs in some respect from other surveys. If, e.g. the net household income was required, only two or three questions might have been sufficient and the questionnaire would be fairly different from that of the EU-SILC.

The features of the measurement of income in the context of the EU-SILC are based on the recommendations of the Expert Group on Household Income Statistics (2001), the so called "Canberra Group" and the ESSPROSS manual for social transfers received which have been adopted in the target variables. The measurement itself has to rely on the more differentiated national categories, which could in principle be grouped according to the international conventions without major problems. In order to assure precise income information on household level, the person who knows best about household affairs shall

answer the questionnaire. This rule shall help to get the best income information from each private household. For personal income information, according to the framework regulation of EU-SILC, the mode of data collection is a personal interview. This will be realised in the German case by the use of separate self-administered personal surveys for every respondent in order to prevent an underestimation of the personal income.

However, for the international harmonisation there are numerous difficulties which arise from differences in the national systems of income and social security. Just to name a few examples: In some countries it is not usual for employees to get information on their gross income. Sick leave payment might be part of the income or of the social security system in different countries. Tax reimbursements have a very different timing and volume in different tax systems. Some income components might fit in different income categories according to the national structures (e.g. subsidies for housing cost, meals, transport which can be included in the income or paid separately). Finally, the retrospective income measurement is a further source of error. Respondents have to recall the income from the last year as for many income components no information available before the spring of the following year (or even later).

4 Conclusion

Realising cross-national harmonisation is a challenging process for statistical agencies in which no ideal solution is at sight. Instead, feasible solutions have to be looked for in some kind of an optimisation process. Suitable approaches have to take into account the requirements of cross-national and cross-survey harmonisation within the national statistical systems as well as the restrictions related to possible survey designs and data collection modes. Although the specific situation of each variable has to be analysed individually, some general conclusions can be drawn.

From the point of view of cross-national harmonisation, theoretically, one could think of input harmonisation as an ideal solution. At first sight, harmonising all concepts and methods seems to the best way to obtain harmonised results. The problem is that while the survey itself can be harmonised to a large extent, the national and cultural environment of the survey remains quite heterogeneous which leads to a number of problems. Problems arise from the fact that most variables cannot be removed from their national setting (e.g. the income variables referred to above). Other harmonised variables cannot be measured appropriately as respondents do not have the international concepts available in their everyday knowledge.

With regard to the requirements of the users of the national-level data, output harmonisation seems to be the preferable strategy. Input harmonised surveys often lack a foundation within the respective national statistical systems. This potentially leads to survey results which could be used only partially on the national level (as shown in the case of occupation variables). Further possible consequences include the longer time span required for data processing, increased survey costs, and lacks of coherence with other statistics from the same socio-economic area. Numerous problems of this kind have been experienced in the context of the European Community Household Panel (ECHP).

However, ex-post output harmonisation cannot be regarded as the best solution either. Although offering the possibility to keep the national statistical systems unchanged (and thus avoiding breaks in time series etc.) the extensive research carried out in the context of CHINTEX as well as the German pilot study for the implementation of the EU-SILC (in which the data provision from an existing national data source also was tested) made clear that ex-post output harmonisation could not be applied as a standard procedure. As CHINTEX showed (Minkel, 2004), ex-post output harmonisation is subject to an important number of preconditions which are given only in a fairly small number of cases and only for a restricted set of survey variables.

Against this background, ex-ante output harmonisation seems to be the most promising way to enable international harmonisation, at least in the case of EU-SILC. This strategy gives flexibility to use the "best national practices" and to reconcile national and international information requirements. Complete harmonisation probably is only a theoretical option.

References

Commission of the European Communities. 2003. *Joint Report on Social Inclusion Summarising the Results of the Examination of the National Action Plans for Social Inclusion (2003-2005).* Brussels: Commission of the European Communities.

Ehling, M., U. Rendtel, et al. 2004. *Harmonisation of Panel Surveys and Data Quality.* Wiesbaden: Statistisches Bundesamt.

Emmerling, D., and T. Riede. 1997. „40 Jahre Mikrozensus." *Wirtschaft und Statistik* 3:160-174.

Eurostat. 2000. Working Group Document E0/00/DSS/2/6/EN. *Harmonisation of Core Variables.* Luxembourg: Eurostat.

Eurostat. 2001. Working Group Document EU-SILC 49/01. *Concepts.* Luxembourg: Eurostat.

Eurostat (Eds.). 2002. *Quality in the European Statistical System – the Way Forward.* Luxembourg: Eurostat.

Eurostat (Eds.). 2003. *Household Budget Surveys in the EU.* Luxembourg: Office for Official Publications of the European Communities.

Expert Group on Household Income Statistics, The Canberra Group (Eds.). 2001. *Final Report and Recommendations.* Ottawa.

Geis, A., and J. Hoffmeyer-Zlotnik. 2001. „Kompatibilität von ISCO-68, ISCO-88 und KldB-92." *ZUMA-Nachrichten* 48:117-138.

Grais, B. 1999. "Statistical Harmonisation and Quality." Pp. 51-114 in *The Future of European Social Statistics,* edited by Eurostat. Luxembourg: Eurostat.

Günther, R. 2004. "Experiences from ECHP Conversions." Pp. 8-38 in *Harmonisation of Panel Surveys and Data Quality,* edited by Ehling, M., U. Rendtel et al. Wiesbaden: Statistisches Bundesamt.

Hoffmeyer-Zlotnik, J. H.P., D. Hess, and A. Geis. 2004. „Computeruntzerstützte Vercodung der International Standard Classification of Occupation (ISCO-88)." *ZUMA-Nachrichten* 55:29-52.

Hoffmeyer-Zlotnik, J. H.P., and C. Wolf. 2004. Paper presented at the European Conference on Quality and Methodology in Official Statistics (Q2004), Mainz, Germany 24-26 May 2004. *Comparing Demographic Variables Across Nations: From Concept to Indicator.*

International Labour Office (Eds.). 1990. *ISCO-88. International Standard classification of Occupations.* Geneva: ILO.

Mejer, L. 2003. "Harmonisation of Socio-Economic Variables in EU Statistics." Pp. 67-85 in *Advances in Cross-National Comparison,* edited by Hoffmeyer-Zlotnik, J. H.P., and C. Wolf. New York etc.: Kluwer Academic/Plenum Publishers.

Meyer, I. 2004. Paper presented at the European Conference on Quality and Methodology in Official Statistics (Q2004), Mainz, Germany 24-26 May 2004. *Strategies in Dealing with New Data Needs and their Consequences on Quality.*

Minkel, H. 2004. "Methodology of Panel Data Conversion Using Additional Panel and Cross-survey Information." Pp.39-64 in *Harmonisation of Panel Surveys and Data Quality,* edited by Ehling, M., U. Rendtel et al. Wiesbaden: Statistisches Bundesamt.

Schwarz, N. 2001. "The German Microcensus." *Schmollers Jahrbuch* 121:649-654.

Statistisches Bundesamt (Eds.). 1992. *Klassifizierung der Berufe.* Wiesbaden: Statistisches Bundesamt.

Timm, U. 2004. Paper presented at the European Conference on Quality and Methodology in Official Statistics (Q2004), Mainz, Germany 24-26 May 2004. *The Impact of Questionnaire Design on Item Nonresponse Concerning Income Questions.*

United Nations Educational, scientific and cultural Organization (Eds.). 1997. *International Standard Classification of Education. ISCED 1997.* Paris: Unesco.

HARMONISING BACKGROUND VARIABLES IN THE EUROPEAN SOCIAL SURVEY

KIRSTINE KOLSRUD & KNUT KALGRAFF SKJÅK

There is no straightforward solution to the comparative measurement of demographic and socio-economic variables in international surveys. The paper will present the ESS approach, where the background variables have been developed by a centrally coordinated team of experts and afterwards clearly defined in a source questionnaire. The advantages and disadvantages of this approach are discussed. The paper ends with a closer look at the coding of Educational level in the ESS, and points out critical considerations for a successful harmonisation.

1 The European Social Survey

The European Social Survey (ESS) is an academically driven social survey with the central aim of developing and conducting a systematic study of changing values, attitudes, attributes and behaviour patterns within European polities. At the same time the ESS aims to raise the methodological standards to which cross-national research is carried out. The data collection is planned to be carried out every two years, by means of face-to-face interviews of about an hour in duration, followed by a short self-completion supplement. The questionnaire consists of a "core" module lasting for about half an hour, and will remain relatively constant form round to round. I addition there are two or three "rotating modules" repeated at intervals, each of which is devoted to a substantive topic or theme. The modules are selected following a Europe-wide competition. Thus, while the purpose of the rotating modules is to provide an in-depth focus on a series of particular academic or policy concerns, the core module aims instead to monitor change and continuity in a wide range of socio-economic, socio-political, socio-psychological and socio-demographic variables, and to provide background variables for the analysis of the rotating modules. (ESS Questionnaire Development Report *http://naticent02.uuhost.uk.uu.net/questionnaire/que_development_report.htm)*

The first data and documentation from ESS 2002/2003, is already freely available to social researchers in Europe and beyond from *http://ess.nsd.uib.no*. Data collection for the second round of the ESS will be staring in September 2004, and data will be available in September 2005.

1.2 The background of the European Social Survey

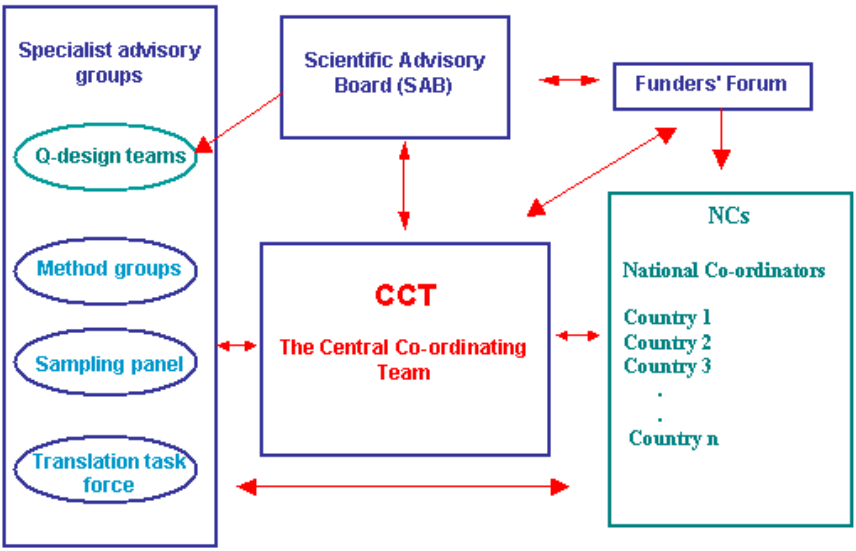
The idea for a European Social Survey originated from the experience in the collaborative comparative research project "Beliefs in Government" funded by the European Science Foundation (ESF) in the early Nineties (Kaase, 2003). The project work of "Beliefs in Government" made it apparent that there was little in terms of national survey evidence, which could meet the minimal criterion of at least functional equivalence across countries (ESF, 1999)

On the initiative of the Standing Committee of the Social Sciences (SCSS) of the ESF, an expert group (led by Max Kaase, one of the two co-directors of the Beliefs in Government project, and member of the SCSS) was set up to develop some criteria for an eventual ESS. Their work led to the SCSS creating a Steering Committee, composed of national representatives nominated by ESF member organisations, and a Methodology Committee with a group of experts to plan the survey in a most meticulous way. In 1999 the SCSS gave their approval green light to go ahead with the implementation of ESS (Kaase, 2003).

Not only is the ESS concept and methodology developed and planned over several years by leading scholars in their field, the funding structure of the ESS does also represent an innovation. In close collaboration with the "Research Directorate-General" of the European Union and ESF, a funding scheme was set up whereby the research councils of the ESF member organisations in the participating countries would cover the cost for the national surveys, and a Central Co-ordinating Team would be funded based on the competitive proposal for EU funding (Kaase, 2003). The central co-ordination of the ESS has so far, in competition with other research projects, been funded by the European Commission's Fifth Framework Programme for Round 1 (ESS, 2002/2003) and for Round 2 (ESS, 2004/2005), and by the Sixth Framework Programme for Round 3 (ESS, 2006).

Partly due to the financing structure and partly due to the aims of raising the methodological standards by which cross-national research is carried out, the organisational structure of the ESS has developed into a rather complex structure of a Central Co-ordinating Team (CCT) surrounded and supported by a number of expert groups and advisory bodies, see Figure 1 below.

Figure 1 Organisational Structure of the ESS



The funding of a Central Co-ordinating Team gives the ESS a rather unique opportunity to develop common methodological standards, make specifications, protocols, and in general guide and assist the participating countries thus enhancing the harmonisation and the comparability of the survey. The degree of standardisation and monitoring from the CCT is in fact one of the features that sets the ESS apart from other cross-national surveys (Bryson & O'Shea, 2003)

The organisational structure of the ESS does, however, not only reflect a top-down approach, but does clearly have bottom-up elements whereby questionnaire design groups and researchers from the participating countries can play a central role in designing the questionnaire and determining how the project develops (the Questionnaire design teams and the National co-ordinators, in green). The aim is that views and needs of those implementing the survey in each country (the national co-ordinators) can be taken into account when the centrally-designed protocols and questionnaires are drawn up. After all what the ESS (and other cross-national surveys) strive for is to balance the desired level

of comparability while at the same time being appropriately sensitive and responsive to cultural differences. At the same time, one of the major challenges for the ESS in building a time series is to ensure consistency and standardisation whilst ensuring the highest methodological standards (Bryson & O'Shea, 2003).

As already mentioned the centrally funded co-ordination of the ESS provides the CCT with ample opportunity to develop common standards and protocols. This top-down feature of the ESS does (in contrast to for example the ISSP) at least give the ESS the opportunity to apply input harmonisation of all variables, including background variables.

1.2 Development of the ESS core questionnaire

Even tough the development of the ESS core questionnaire is the responsibility of the CCT, it has none the less been constructed with help and guidance from advisory groups like the Scientific Advisory Board, National co-ordinators and other commentators. The design of the core questionnaire started early on when the former Steering and Methodology Committees determined their priorities for topics to be included within the core. Expert papers were commissioned to provide both a substantive overview of the concepts in each selected field and, where possible, a set of recommended questions that would successfully tap these concepts cross-nationally.

The questionnaire design process can be described in 6 stages:

Stage1
The first task was to ensure that the various concepts that were to be included (based on the expert papers) were actually represented as precisely as possible by the candidate questions and response scales.

Stage 2
To achieve the appropriate quality standard, the questions and scales, wherever possible, underwent an evaluation using standard quality criteria such as reliability and validity. These evaluations were carried out by Willem Saris and his colleagues, using the program SQP developed for the prediction of the reliability and validity of questions on the basis of more than 1000 questions evaluated by MTMM studies (Scherpenzeel & Saris, 1997; Saris et al., 2003). Attention was also given to other considerations such as scalability and internal consistency, comparability of items over time and space, expected item non-response, social desirability and other potential biases, and the avoidance of ambiguity, vagueness and double-barrelled questions.

Stage 3

The next step was the first translation from the source language (English) into one other language for the purpose of two large-scale national pilots. The translation panel, which is convened by Janet Harkness at ZUMA, guided this process to ensure optimal comparability between the two versions.

Stage 4

The fourth step was the two-nation pilot itself, which also contained a number of split-run experiments on question wording alternatives. Most of these experiments were in a drop-off self-completion supplement, but some were in the main interview questionnaire.

Stage 5

The pilot was analysed in detail to assess both the quality of the questions and the distribution of the substantive answers. Problematical questions, whether on grounds of weak reliability or validity, or because they turned out to produce deviant distributions or weak scales, were sent back to the drawing board.

Stage 6

The final step was the production of a fully-fledged 'source questionnaire', ready for translation from English into all ESS languages. The ESS aim was to apply a sequential Ask-the Same-Questions model – one in which the source questionnaire is finalised first and the translations produced. The English source questionnaire was annotated to aid the translation process. This annotation was carried out in collaboration with the various question authors and National co-ordinators in order to avoid ambiguities by providing definition and clarifications of the concept behind questions, especially where the words themselves were unlikely to have direct equivalents in other languages. Each participating country then carried out a small-scale pre-test to iron out any remaining translation or substantive issues.

(ESS Questionnaire Development Report, *http://naticent02.uuhost.uk.uu.net/questionnaire/que_development_report.htm)*

The result of the ESS questionnaire development is a carefully centrally designed questionnaire with valuable input and guidance from a number of advisory expert groups and researchers.

1.3 Background variables in the ESS

Although there is an increasing need for the comparative measurement of demographic and socio economic variables, no such comprehensive standards for social research exist at the European level (Wolf & Hoffmeyer-Zlotnik, 2003). No matter how attractive the idea, the ESS did hence not have the option of adopting an existing (and well tested) set of background measures.

The development of the background variables in the ESS did in fact arise from a similar process as the rest of the core questionnaire, the key specialist paper being written by the scholars Robert Erikson & Jan O. Jonsson. The resulting topics for the ESS demographic and socio-economic variables are listed below. For a complete list of the demographic and socio-economic variables in ESS Round 1, please see Appendix.

- Demography
- Family, household
- Education
- Employment, main activity
- Economic standing, income
- Heritage, identity, religion

The topics were transformed into constructs, questions and answer scales. However, it is beyond the scope of this paper to go into detail on how all these topics were transformed into constructs and questions. We have instead selected three variables, which we would like to use as examples of the kind of considerations that were undertaken by the CCT when deciding on how to measure the constructs.

Before we go into some detail on the selected variables: Occupation, Religion and Education we would like to make a more general comment on the use of international coding frames in the ESS.

1.4 The use of international standards for coding of background variables

Even tough there are no common set of measures for background variables, there are instruments for measurement of single variables that are established in internationally research (Wolf & Hoffmeyer-Zlotnik, 2003). Naturally the ESS would like to base the relevant background variables upon already existing and accredited standards. The post coding of variables like occupation, industry, language and country were hence decided to be coded into well accredited international standards like the ISCO-88 (com) for "Occupation", NACE rev1. for "Industry" and ISO 3166-1 and ISO 639-2 for the coding of country and language respectively. All established standards having been developed by, or in close cooperation with, international organisations such as the International Labour Office, the United Nations and Eurostat.

For other variables, like education and religion the decision was not so clear-cut as to whether the coding should be according to an existing standard or whether the ESS should develop their own coding scheme, or simply leave some of the variables country specific.

Occupation

Based on the expert papers as well as an investigation into the use of coding standards for occupation in other cross-national surveys (the International Social Survey Programme, the European Community Household Panel and others) it seemed quite evident that the International Standard Classification of Occupation, ISCO-88 would be the natural choice for the ESS. Not only could the standard coded at 4 digit level (with the addition of information on standing in employment) provide sufficient information to construct measures of both social class and occupational prestige, but the use of skill level as one of the "aggregation levels" in the standard does also provide the users with a direct indicator of labour market position. (Erikson & Jonssson, 1999).

Table 1 Occupation, ESS

Construct	Occupation
ESS Questions	• Verbatim recorded questions: • What is/was the name or title of your main job? (F21) • In your main job, what kind of work do/did you do most of the time? (F22) • What training or qualifications are/were needed for the job? (F23) • What does/did the firm/organisation you work/worked for mainly make or do? (F24) • Including yourself, about how many people are/were employed at the place where you usually work/worked? (F15)
Standard/coding frame	• ISCO88 (com) 4 digit
Harmonisation, comments	• Mainly input-harmonisation, but the participating countries could ask additional questions if further information was considered necessary for the coding of the standard.
Problems	• The knowledge and acquaintance with the ISCO-88 (com) coding framework varied between the countries. Some countries coded into their national standard for occupation and bridged that coding into ISCO-88 (com), others coded into the ISCO-88 ILO version of the standard and adapted the coding to ISCO-88 (com) afterwards. Others coded directly into the ISCO-88 (com) framework. • The different approaches to the coding can all yield different kind of errors, none of which are detectable for the ESS Archive after the coding is done. A common problem with post coded variables. The bridging from national standards to the ISCO8-8 (com) is not documented in the survey.

It was, however, decided by the CCT to use what can best be described as the European Union variant of the ISCO-88, namely the ISCO-88 (com). According to Eurostat, the ISCO-88 (com) should be used for EU purposes. The ISCO-88 (com) should not be regarded as a different classification from ISCO-88, but is the result of a coordinated effort by National Statistical Institutes to implement ISCO-88 for census and survey coding purposes within the European Union (Elias & Birch, 1994).

Religion

While it is relatively straightforward to apply international standards for the coding of labour market characteristics like occupation, it is a much greater challenge to find common denominators for the coding of variables for religion and education (Erikson & Jonsson, 1999).

The problems of finding a common measure for religion can undoubtedly be traced back to the differences in the historical developments of religious institutions across countries. Different colonial and immigration history leading to different religious minorities, different historical developments in the secular realm and the presence of state churches or not, all contribute to a very diverse religious landscape, even within Europe (Hoffmeyer-Zlotnik & Wolf, 2003).

Recognising the need for national expertise in this field, the CCT settled for a combination of input and output harmonisation for the religious variables in the ESS, see Table 2. Although the same question(s) were asked in all countries, the answer categories (denominations listed) should be set up by the national teams to best match the religious landscape of their country, and later re-coded into the common ESS coding frame.

Education

Educational systems differ markedly across countries. They have been formed by nation specific cultural and social concepts and traditions, and depend on national regulations and legislations (Hoffmeyer-Zlotnik & Wolf, 2003). In addition educational systems are also more often subject to change than occupations and economic activities. Hence, harmonising and comparing variables measuring different aspects of education is rather difficult. Not only does comparison and harmonisation require substantial knowledge on the various national structures, but ideally the changes over time should also be taken into consideration (Braun & Mohler, 2002).

There are some simple ways of identifying a common educational structure (such as primary, secondary tertiary levels etc.), but there has also been developed more ambitious coding schemes like the CASMIN schema developed by Walter Müller and associates (Müller & Shavit, 1998; in Erikson & Jonsson, 1999), and the UNESCO international standard classification of education (ISCED).

Table 2 Religion, ESS

Construct	Religion
ESS Questions	• Do you consider yourself as belonging to any particular religion or denomination? (C9) • Which one? (C10) Country specific answer categories.
Standard/coding frame	• ESS coding frame. 1 Roman Catholic 2 Protestant 3 Eastern Orthodox 4 Other Christian Denominations 5 Jewish 6 Islam 7 Eastern Religions (Buddhist, Hindu, Sikh, Shinto, Tao etc.) 8 Other Non-Christian Religions
Harmonisation, comments	• The same two questions were to be asked in all participating countries, but the answer categories were set up by the national teams to suit the needs of their country. The responses were then to be post-coded into a common ESS coding frame for religious denominations. • Elements of both input harmonisation (same question) and output harmonisation (different answer categories bridged to common standard)
Problems	• Ensuring that all the national answer categories was suited for bridging into the common ESS coding framework (exhaustive and mutually exclusive). • Differences in national structures of religious organisations (state church – no state church) • In some countries (with a state church) there is at least a potential risk of measuring membership, rather than affiliation.

After careful consideration the CCT settled for the use of a slightly modified ISCED 1997 coding frame for the coding of highest level of education in the ESS. The coding frame is listed in Table 3 below. The coding frame does only distinguish between the main levels of the ISCED 1997. The sub-level information on educational direction is not included. The first category 0 "Pre-primary education" has also been changed to "Not completed primary (compulsory) education".

As a supplement to this rather crude coding frame it was decided that the participating countries should also have the opportunity to include the country specific variable(s) which were bridged into the ESS coding frame. Documentation of the bridging into the common coding frame was a requirement. Hence the choice of the ESS for the coding of education was both a country specific non-harmonised variable and an output harmonisation variable.

Table 3 Education, ESS

Construct	Education (highest level of education)
ESS Questions	• Country specific question(s)
Standard/coding frame	• Country specific coding frames + • ESS coding frame (modified ISCED 1997) 0 Not completed primary (compulsory) education 1 Primary or first stage of basic 2 Lower secondary or second stage of basic 3 Upper secondary 4 Post secondary, non-tertiary 5 First stage of tertiary 6 Second stage of tertiary
Harmonisation, comments	• A case of both no-harmonisation (including the country specific variables in the data files) and output harmonisation (re-coding into an ESS specific coding frame for educational level).
Problems/advantages	• Ensuring that the national answer categories are such that they best can be re-coded into the ESS common coding frame. • The advantage being both providing educational experts with the nation specific as well as having a common standard re-coded by the national teams. • Requires thorough documentation of the national educational system and how the country specific variables is bridged into the ESS coding frame.

2 Case: Output Harmonisation of Education Level in the ESS

In this last section we will present some findings indicating how a common coding frame for education taps the construct, compared to the country-specific variables. The first observation is that when output harmonisation results in a higher level of aggregation of sub-groups (because a large number of country-specific categories have to be collapsed into more general concepts) important characteristics of the data might be lost. The second observation is that national teams don't necessarily have a shared understanding of the product (common frame) variable, even when the categories of the product variable are well defined.

Table 4 Bivariate Analysis of Dependent Variable "Attitudes towards Immigration" Index and Education Levels (Regressors). Reference Group: Primary or First Stage of Basic Education. Parameters Estimates of ESS Coding frame Categories (Netherlands)

	DF	Parameter estimates	Standard error	t value	Pr > |t|	N
Intercept	1	16.86802	0.37977	44.32	<.0001	197
0. Not completed primary education	1	0.27484	1.47435	0.25	0.8521	14
2. Lower secondary or second stage of basic	1	*0.48850*	*0.43059*	*1.34*	*0.2567*	*690*
3. Upper secondary	1	2.03294	0.43545	4.88	<.0001	626
4. Post secondary, non-tertiary	1	2.97256	0.59171	5.18	<.0001	138
5. First stage of tertiary	1	3.49401	0.44702	8.03	<.0001	511
6. Second stage of tertiary	1	3.57642	1.81692	2.01	0.0491	9

Aggregation

Tables 4 and 5 below show the result of bivariate regression where an index of attitudes towards immigration is the dependent variable and education levels are regressors. The index is an additive scale of four variables from the following questions in the ESS 2002/2003 data file, and we have used data from the Netherlands as example:

D25 Using this card, would you say that people who come to live here generally take jobs away from workers in [country], or generally help to create new jobs?

D28 And, using this card, would you say that [country]'s cultural life is generally undermined or enriched by people coming to live here from other countries?

D29 Is [country] made a worse or a better place to live by people coming to live here from other countries? Please use this card.

D30 Are [country]'s crime problems made worse or better by people coming to live here from other countries? Please use this card.

All questions have an eleven-point scale, where 0 indicates negative consequences of immigration, 10 positive consequences. The scale has proved to have high validity across countries (Billiet, 2003).

From a data-explorative point of view, one obvious result in Table 4 is the low significance (t=1.34, N=690) of the parameter estimate of the sub-group 2. "Lower secondary or second stage of basic". The result might indicate that the group is very heterogeneous with respect to attitudes toward immigration, compared to other groups.

Table 5 Bivariate Analysis of Dependent Variable "Attitudes towards Immigration" Index and Education Levels (Regressors). Reference Group: Primary or First Stage of Basic Education. Parameters Estimates of Country Specific Categories (The Netherlands)

	DF	Parameter estimates	Standard error	t value	Pr > |t|	N
Intercept	1	16.86802	0.37895	44.51	<.0001	197
Not completed primary school	1	0.27484	1.47115	0.19	0.8518	14
Lower secondary school, technical (lbo)	1	*0.15228*	*0.46411*	*0.33*	*0.7429*	*394*
Lower secondary school, theoretical (mulo,mavo)	1	*0.93603*	*0.48905*	*1.91*	*0.0558*	*296*
Short upper sec. professional (kmbo, vhbo)	1	0.72657	0.95299	0.76	0.4459	37
Upper secondary professional education (mbo)	1	1.69289	0.46411	3.65	0.0003	394
Higher secondary school (mms, havo)	1	2.97256	0.59042	5.03	<.0001	138
Pre-scientific secondary school (hbs, vwo)	1	3.05751	0.66675	4.59	<.0001	94
Post secondary, non-tertiary education (mbo plus)	1	2.88445	0.65092	4.43	<.0001	101
Tertiary professional education (hbo)	1	3.28987	0.47113	6.98	<.0001	361
Tertiary scientific education, university	1	3.93043	0.60241	6.52	<.0001	129
Tertiary post-scientific education (teachers, doctors)	1	4.32246	1.22095	3.54	0.0004	21
Second stage of tertiary education, Ph.D. education	1	3.57642	1.81297	1.97	0.0487	9

The same analysis using the country-specific education variable unfolds that the harmonised education variable covers up significant differences in the "Lower secondary or second stage of basic" category (Table 5). The group with theoretical education from the lower secondary level seems to score much higher (more positive attitudes) on the immigration scale than people with technical education, a result that might suggest the importance of for example occupation and life career in forming attitudes towards immigration. This is an example of how important it is to carefully consider the need for country-specific variables in addition to a common coding frame.

2.2 Reliability of bridging

Successful output harmonisation depends on reliable bridging of the country-specific measurements into the common coding frame. As described in section 1.4, the ESS coding frame for education is a slightly modified version of UNESCO's ISCED-1997 classification. In the ESS Round 1, UNESCO's Operational Manual (UNESCO, 1999) was made available to the national teams together with the coding frame to ensure a *shared understanding* of how to apply the ISCED to the their national data.

Looking at Table 6 on page 177, we observe that the three countries we have used as examples have bridged their country-specific education variables in different ways. Some

of the differences might reflect the educational systems, for example the length of primary or compulsory education. A general "problem" in harmonising education in the ESS countries seems to be to draw comparable borders between the primary and secondary levels. Table 6 also documents large differences in the educational system and/or coding practices with regard to the ESS category 4, "Post secondary, non-tertiary".

However, the coding of the ESS categories 5 and 6 obviously reflects different understandings of the product variable. The UNESCO's Operational Manual is quite clear in its distinction between the two categories:

ISCED level 5 – First stage of tertiary education (not leading directly to an advanced research qualification):

"Qualifications acquired at the end – The programmes at Level 5 do not lead directly to an advanced research degree (which is actually an ISCED level 6 qualification), but to other degrees or diplomas which may or may not have a research component. In most countries, some programmes at Level 5 lead to a first university degree (a Bachelor's degree or its equivalent), and others lead to a second more advanced degree (a Master's degree or its equivalent). Both of these are to be classified at Level 5. In some countries, there is only one long-duration programme that leads to a degree that is equivalent to the combined Bachelor's and Master's degree programmes in other countries." (UNESCO, 1999: 31).

ISCED Level 6 – Second stage of Tertiary Education (leading to an Advanced Research Qualification):

"Destination of the Graduates – Those who successfully complete the programmes of Level 6 are generally eligible for faculty positions in universities and other institutions offering ISCED 5A programmes, as well as for research posts in government, industry, and other organisations employing researchers." (UNESCO, 1999: 37).

The main criterion for level 6 is that it is reserved for programmes that lead directly to an advanced research qualification. This criterion might very well be ambiguous in some educational systems, but when taking the cited main feature for level 5 into consideration, it seems clear that only degrees above a "Bachelor's degree or its equivalent" or a "Master's degree or its equivalent" should be coded into category 6 in the ESS coding frame.

While a majority of the ESS countries have coded only the Ph.D. degree (or its equivalent) into category 6, the bridges documented in Table 6 are all examples of the inclusion of lower degrees. In Belgium it is possible to move groups from category 6 to 5, while the country-specific variables of Israel and Sweden do not fully distinguish between levels

corresponding to Bachelor, Master and Ph.D. It is therefore not possible to bridge these country-specific variables into the common coding frame.

As mentioned in section 1.4, educational systems differ markedly across countries, and education is certainly one of the most difficult measures to harmonise in cross-national surveys. The more ambitious coding schemes like CASMIN and ISCED are certainly powerful tools for measuring and coding education, making us able to code education at a detailed and comparable level. However, detail has to be balanced against several other considerations. Firstly, collecting very detailed information on education would occupy a large amount of interviewing time, expelling other questions. Secondly, the more detailed the information and coding is, the less significant will the sample size of each category be. And thirdly, most users of the data would not like to face the burden of organising the detailed information into more comprehensible patterns.

In our view, the ESS has been successful in defining its set of comparable background variables and coding frames by giving ample opportunity to capture national variations by use of country-specific coding and variables, and at the same facilitating cross-national comparison by use of standards and standardised variables. Rather than using resources in searching or developing new standards, we believe that the largest potential for improving the background variables in the ESS can be found in better working procedures. Still with education as an example, the quality of measurement would be greatly improved by better training of interviewers and/or keying personnel in their country's educational system, and as the example in Table 6 clearly indicates, a centrally coordinated review of the national input instruments and the national bridging into the common ESS coding frames would have to be considered.

Table 6 Bridging of Country Specific Education Categories to ESS Coding Frame. Israel, Belgium and Sweden

ESS coding frame	Israel	%	ESS %	Belgium	%	ESS %	Sweden	%	ESS %
0. Not completed primary education	No formal qualification	1.9	1.9	Niet voltooid lager onderwijs	1.6	1.6	Not finished elementary school	1.3	1.3
1. Primary or first stage of basic	Lowest formal qualification	8.4	8.4	Lager onderwijs, basisschool	13.5	13.5	Elementary school, old Elementary school	15.4 11.7	27.1
2. Lower secondary or second stage of basic	Not finish vocational high school Not finish general high school	6.1 7.5	13.6	Lager beroepsonderwijs Lager technisch onderwijs Lager algemeen secundair onderwijs	10.4 0.0 9.5	19.9	Lower secondary and elementary school, old Vocational school 1963-1970 2 year high school	3.8 2.2 13.5	19.5
3. Upper secondary	Full voc hs without matriculation certificate Full voc hs with matriculation certificate Full general hs without matriculation certificate Full general hs with matriculation certificate Yeshiva hs without full matriculation certificate Yeshiva hs with full matriculation certificate	7.2 6.0 6.3 11.8 0.5 0.9	32.7	Hoger secundair beroepsonderwijs Hoger technisch onderwijs Hoger algemeen secundair onderwijs	10.4 13.5 13.5	37.4	3-4 year high school prior 1995 Vocational high school after 1992 Theoretical high school after 1992	13.5 4.9 3.2	21.6
4. Post secondary, non-tertiary	Post secondary	11.5	11.5	Hoger onderwijs van het korte type	14.0	14.0			
5. First stage of tertiary	Not finish University degree University Ba degree completed	8.1 14.6	22.7				University, no exam University, exam less than 3 years	8.6 6.0	14.6
6. Second stage of tertiary	University Ma /Phd degree completed	9.2	9.2	Hoger honderwijs van het lange type Universiteit Doctoraat	4.2 7.7 1.7	13.6	University, exam more than 3 years	16.1	16.1

3 Conclusion

Background variables "allow us to define contexts in which respondents' opinions, attitudes and behaviour are socio-economically embedded" (Braun & Mohler, 2002: 112). The measurement of background variables and definitions of homogenous sub-groups in mono-cultural or national surveys is based on knowledge of national concepts, rules and structures, and in cross-national surveys these cultural-specific measurements have to be harmonised into equivalent measures (Hoffmeyer-Zlotnik & Wolf, 2003).

The central funding and the organisational structure has given the ESS a unique opportunity to develop such a set of equivalent measures, mainly based on input harmonisation and internationally accredited standards. In Round 1 of the ESS, the project has achieved a lot in balancing a high level of comparability while at the same time being appropriately sensitive and responsive to cultural differences. The ESS will also provide a basis for further improvements in the harmonisation of cross-national measures, as feedback from researchers from all over the world will give us knowledge about how the measures have worked.

In the planning years of the ESS the experiences from on-going cross-national surveys like the ISSP were important knowledge bases for considerations and decisions made by the Steering and Methodology Committees (see page 164). When the ESS now has become reality, other international survey projects have the opportunity to utilise the ESS experience in their efforts of improving their background variables.

The background variables in the ESS have been planned in a most meticulous way, but the harmonisation of measurement and coding might still be improved. One example is the measurement of occupation (see page 169), where countries (or their fielding institutes) have long traditions in either using their own standard for occupation, their own variant of the ISCO-88 standard or ISCO-88 ILO rather than ISCO-88 (com). The different approaches to the coding can all yield different kind of errors, none of which might be detectable for the ESS Archive or the data users after the coding is done. The other example is education, where more attention to the country-specific variables as well as the bridging into the coding frame is needed.

Securing that harmonisation results in functional equivalent measures, i.e. they reflect the same phenomenon or dimension is therefore a two-way process:

1. Nationally diverse perspectives have to be taken into account in defining the resulting constructs.
2. National measurements on their side must be adapted to the common coding frames to ensure complete coverage of the constructs.
3. National teams must have a shared understanding of the product of the harmonisation.
4. Centrally co-ordinated assessment of country-specific instruments and bridging before national questionnaires are finalised and signed off to ensure that requirements 1-3 are met.

References

Billiet, J. 2003. *Questions on immigration: selection for the core in second round.* Paper prepared for the CCT. Unpublished.

Braun, M., and P. Ph. Mohler. 2002. "Background Variables." In *Cross-Cultural Survey Methods*, edited by Harkness, A., F. J. R. van de Vijver, and P. Ph. Mohler. New York: Wiley.

Bryson, C., and R. O' Shea. 2003. *The European Social Survey – One survey on 22 countries.* Abstract from ESS Launch Conference in Brussels, November 2003. www.europeansocialsurvey.org.

Elias, P., and M. Birch. 1994. *Establishment of Community-Wide Occupational Statistics. ISCO 88 (COM) A Guide for Users.* Report from University of Warwick, Institute for Employment Research.

Erikson, R., and J. O. Jonsson. 1999. "How to ascertain the socio-structural position of the individual in society. Suggestions for the core ESS questions." In *ESS Questionnaire Development Report.* naticent02.uuhost.uk.uu.net/questionnaire/que_development_report.htm.

ESF. 1999. *The European Social Survey (ESS) – a research instrument for the social sciences in Europe.* Report prepared for the Standing Committee for the Social Sciences (SCSS) of the European Science Foundation (ESF).

Hoffmeyer-Zlotnik, J. H.P., and C. Wolf. 2003. "Comparing Demographic and Socio-Economic Variables Across Nations." Pp. 389-406 in *Advances in Cross National Comparison. A European Working Book for Demographic and Socio-Economic Variables,* edited by Hoffmeyer-Zlotnik, J. H.P., and C. Wolf. New York: Kluwer Academic/Plenum Publishers.

Kaase, M. 2003. *Identifying the need for an ESS.* Abstract from ESS Launch Conference in Brussels, November 2003. www.europeansocialsurvey.org.

Scherpenzeel, A., and W. E. Saris. 1997. "The Validity and Reliability of Survey Questions: a Meta Analysis of MTMM Studies." *Sociological Methods and Practice* 25:341-383.

Saris, W. E., W. van der Veld, and I. Gallhofer. 2003. "Development and improvement of questionnaires using predictions of reliability and validity." In *Questionnaire development, evaluation and testing,* edited by Presser, S., J. Rothgeb, M. Couper, J. Lessler, E. Martin, and E. Singer. New York: Wiley.

UNESCO. 1999. *Operational Manual for ISCED-1997.* (International Standard Classification of Education). UNESCO Institute for Statistics.

Wolf, C., and J. H.P. Hoffmeyer-Zlotnik. 2003. "Measuring Demographic and Socio-Economic Variables in Cross-National Research: An Overview." Pp. 1-13 In *Advances in Cross National Comparison. A European Working Book for Demographic and Socio-Economic Variables,* edited by Hoffmeyer-Zlotnik, J. H. P., and C. Wolf. New York: Kluwer Academic/Plenum Publishers.

Appendix

Demographic and Socio-Economic Variables in the ESS

C9	RLGBLG	"BELONGING TO PARTICULAR RELIGION OR DENOMINATION"
C10	RLGDNM	"RELIGION OR DENOMINATION BELONGING TO AT PRESENT"
C18	CTZCNTR	"CITIZEN OF COUNTRY"
C19	CTZSHIP	"CITIZENSHIP"
C20	BRNCNTR	"BORN IN COUNTRY"
C21	CNTBRTH	"COUNTRY OF BIRTH"
C23	LNGHOMA	"LANGUAGE MOST OFTEN SPOKEN AT HOME: FIRST MENTIONED"
	LNGHOMB	"LANGUAGE MOST OFTEN SPOKEN AT HOME: SECOND MENTIONED"
F1	HHMMB	"NUMBER OF PEOPLE LIVING REGULARLY AS MEMBER OF HOUSEHOLD"
F2	GNDR	"GENDER"
	GNDR2	"GENDER OF SECOND PERSON IN HOUSEHOLD"
	.	
	.	
	.	
	GNDRN	"GENDER OF N'TH PERSON IN HOUSEHOLD"
F3	YRBRN	"YEAR OF BIRTH"
	YRBRN2	"YEAR OF BIRTH OF SECOND PERSON IN HOUSEHOLD"
	.	
	.	
	.	
	YRBRNN	"YEAR OF BIRTH OF N'TH PERSON IN HOUSEHOLD"
F4	RSHIP2	"SECOND PERSON IN HOUSEHOLD: RELATIONSHIP TO RESPONDENT"
	.	
	.	
	.	
	RSHIPN	"N'TH PERSON IN HOUSEHOLD: RELATIONSHIP TO RESPONDENT"
F5	DOMICIL	"DOMICILE, RESPONDENT'S DESCRIPTION"
F6	EDULVL	"HIGHEST LEVEL OF EDUCATION"
	EDLVAT	"HIGHEST LEVEL OF EDUCATION, AUSTRIA"
	EDLVBE	"HIGHEST LEVEL OF EDUCATION, BELGIUM"
	EDLVCH	"HIGHEST LEVEL OF EDUCATION, SWITZERLAND"
	EDLVCZ	"HIGHEST LEVEL OF EDUCATION, CZECH REPUBLIC"
	EDLVDK	"HIGHEST LEVEL OF EDUCATION, DENMARK"
	EDLVES	"HIGHEST LEVEL OF EDUCATION, SPAIN"
	EDLVFR	"HIGHEST LEVEL OF EDUCATION, FRANCE"
	EDLVGB	"HIGHEST LEVEL OF EDUCATION, UNITED KINGDOM"
	EDLVGR	"HIGHEST LEVEL OF EDUCATION, GREECE"
	EDLVHU	"HIGHEST LEVEL OF EDUCATION, HUNGARY"
	EDLVIE	"HIGHEST LEVEL OF EDUCATION, IRELAND"
	EDLVIL	"HIGHEST LEVEL OF EDUCATION, ISRAEL"
	EDLVIT	"HIGHEST LEVEL OF EDUCATION, ITALY"
	EDLVLU	"HIGHEST LEVEL OF EDUCATION, LUXEMBOURG"
	EDLVNL	"HIGHEST LEVEL OF EDUCATION, NETHERLANDS"
	EDLVNO	"HIGHEST LEVEL OF EDUCATION, NORWAY"
	EDLVPL	"HIGHEST LEVEL OF EDUCATION, POLAND"
	EDLVPT	"HIGHEST LEVEL OF EDUCATION, PORTUGAL"
	EDLVSE	"HIGHEST LEVEL OF EDUCATION, SWEDEN"
F7	EDUYRS	"YEARS OF FULL-TIME EDUCATION COMPLETED"
F8a	PDWRK	"DOING LAST 7 DAYS: PAID WORK"
	EDCTN	"DOING LAST 7 DAYS: EDUCATION"
	UEMPLA	"DOING LAST 7 DAYS: UNEMPLOYED, ACTIVELY LOOKING FOR JOB"
	UEMPLI	"DOING LAST 7 DAYS: UNEMPLOYED, NOT ACTIVELY LOOKING FOR JOB"

Appendix (continued)

	DSBLD	"DOING LAST 7 DAYS: PERMANENTLY SICK OR DISABLED"
	RTRD	"DOING LAST 7 DAYS: RETIRED"
	CMSRV	"DOING LAST 7 DAYS: COMMUNITY OR MILITARY SERVICE"
	HSWRK	"DOING LAST 7 DAYS: HOUSEWORK, LOOKING AFTER CHILDREN, OTHERS"
	DNGOTH	"DOING LAST 7 DAYS: OTHER"
	DNGDK	"DOING LAST 7 DAYS: DON'T KNOW"
	DNGREF	"DOING LAST 7 DAYS: REFUSAL"
	DNGNA	"DOING LAST 7 DAYS: NO ANSWER"
F8b	MAINACT	"MAIN ACTIVITY LAST 7 DAYS"
F9	CRPDWK	"CONTROL PAID WORK LAST 7 DAYS"
F10	PDJOBEV	"EVER HAD A PAID JOB"
F11	PDJOBYR	"YEAR LAST IN PAID JOB"
F12	EMPLREL	"EMPLOYMENT RELATION"
F13	EMPLNO	"NUMBER OF EMPLOYEES RESPONDENT HAS"
F14	WRKCTR	"EMPLOYMENT CONTRACT UNLIMITED OR LIMITED DURATION"
	WRKCTRHU	"EMPLOYMENT CONTRACT UNLIMITED OR LIMITED DURATION, HUNGARY"
F15	ESTSZ	"ESTABLISHMENT SIZE"
F16	JBSPV	"RESPONSIBLE FOR SUPERVISING OTHER EMPLOYEES"
F17	NJBSPV	"NUMBER OF PEOPLE RESPONSIBLE FOR IN JOB"
F18	ORGWRK	"TO WHAT EXTENT ORGANISE OWN WORK"
F19	WKHCT	"TOTAL CONTRACTED HOURS PER WEEK IN MAIN JOB OVERTIME EXCLUDED"
F20	WKHTOT	"TOTAL HOURS NORMALLY WORKED PER WEEK IN MAIN JOB OVERTIME INCLUDED"
F21-F23	ISCOCO	"OCCUPATION, ISCO88 (COM)"
F24	NACER1	"INDUSTRY, NACE REV.1"
F25	UEMP3M	"EVER UNEMPLOYED AND SEEKING WORK FOR A PERIOD MORE THAN THREE MONTHS"
F26	UEMP12M	"ANY PERIOD OF UNEMPLOYMENT AND WORK SEEKING LASTED 12 MONTHS OR MORE"
F27	UEMP5YR	"ANY PERIOD OF UNEMPLOYMENT AND WORK SEEKING WITHIN LAST 5 YEARS"
F28	MBTRU	"MEMBER OF TRADE UNION OR SIMILAR ORGANISATION"
F29	HINCSRC	"MAIN SOURCE OF HOUSEHOLD INCOME"
F30	HINCTNT	"HOUSEHOLD'S TOTAL NET INCOME, ALL SOURCES"
F31	HINCFEL	"FEELING ABOUT HOUSEHOLD'S INCOME NOWADAYS "
F32	BRWMNY	"BORROW MONEY TO MAKE ENDS MEET, DIFFICULT OR EASY"
F33	PARTNER	"LIVES WITH HUSBAND/WIFE/PARTNER AT F4"
F34	EDULVLP	"PARTNER'S HIGHEST LEVEL OF EDUCATION"
F35a	PDWRKP	"PARTNER DOING LAST 7 DAYS: PAID WORK"
	EDCTNP	"PARTNER DOING LAST 7 DAYS: EDUCATION"
	UEMPLAP	"PARTNER DOING LAST 7 DAYS: UNEMPLOYED, ACTIVELY LOOKING FOR JOB"
	UEMPLIP	"PARTNER DOING LAST 7 DAYS: UNEMPLOYED, NOT ACTIVELY LOOKING FOR JOB"
	DSBLDP	"PARTNER DOING LAST 7 DAYS: PERMANENTLY SICK OR DISABLED"
	RTRDP	"PARTNER DOING LAST 7 DAYS: RETIRED"
	CMSRVP	"PARTNER DOING LAST 7 DAYS: COMMUNITY OR MILITARY SERVICE"
	HSWRKP	"PARTNER DOING LAST 7 DAYS: HOUSEWORK, LOOKING AFTER CHILDREN, OTHERS"
	DNGOTHP	"PARTNER DOING LAST 7 DAYS: OTHER"
	DNGDKP	"PARTNER DOING LAST 7 DAYS: DON'T KNOW"
	DNGNAPP	"PARTNER DOING LAST 7 DAYS: NOT APPLICABLE"
	DNGREFP	"PARTNER DOING LAST 7 DAYS: REFUSAL"
	DNGNAP	"PARTNER DOING LAST 7 DAYS: NO ANSWER"

Appendix (concluded)

F35b	MNACTP	"PARTNER'S MAIN ACTIVITY LAST 7 DAYS"
F36	CRPDWKP	"PARTNER, CONTROL PAID WORK LAST 7 DAYS"
F37-F39	ISCOCOP	"OCCUPATION PARTNER, ISCO88 (COM)"
F40	EMPRELP	"PARTNER'S EMPLOYMENT RELATION"
F41	EMPLNOP	"NUMBER OF EMPLOYEES PARTNER HAS"
F42	JBSPVP	"PARTNER RESPONSIBLE FOR SUPERVISING OTHER EMPLOYEES"
F43	NJBSPVP	"NUMBER OF PEOPLE PARTNER RESPONSIBLE FOR IN JOB"
F44	WKHTOTP	"HOURS NORMALLY WORKED A WEEK IN MAIN JOB OVERTIME INCLUDED, PARTNER"
F45	EDULVLF	"FATHER'S HIGHEST LEVEL OF EDUCATION"
F46	EMPRF14	"FATHER'S EMPLOYMENT STATUS WHEN RESPONDENT 14"
F47	EMPLNOF	"NUMBER OF EMPLOYEES FATHER HAD"
F48	JBSPVF	"FATHER RESPONSIBLE FOR SUPERVISING OTHER EMPLOYEES"
F50	OCCF14	"FATHER'S OCCUPATION WHEN RESPONDENT 14"
	OCCF14IE	"FATHER'S OCCUPATION WHEN RESPONDENT 14, IRELAND"
F51	EDULVLM	"MOTHER'S HIGHEST LEVEL OF EDUCATION"
F52	EMPRM14	"MOTHER'S EMPLOYMENT STATUS WHEN RESPONDENT 14"
F53	EMPLNOM	"NUMBER OF EMPLOYEES MOTHER HAD"
F54	JBSPVM	"MOTHER RESPONSIBLE FOR SUPERVISING OTHER EMPLOYEES"
F56	OCCM14	"MOTHER'S OCCUPATION WHEN RESPONDENT 14"
	OCCM14IE	"MOTHER'S OCCUPATION WHEN RESPONDENT 14, IRELAND"
F57	ATNCRSE	"IMPROVE KNOWLEDGE/SKILLS: COURSE/LECTURE/CONFERENCE, LAST 12 MONTHS"
F58	MARITAL	"LEGAL MARITAL STATUS"
F59	LVGHW	"CURRENTLY LIVING WITH HUSBAND/WIFE"
F60	LVGOPTN	"CURRENTLY LIVING WITH ANOTHER PARTNER THAN HUSBAND/WIFE"
F61	LVGPTN	"CURRENTLY LIVING WITH PARTNER"
F62	LVGPTNE	"EVER LIVED WITH A PARTNER WITHOUT BEING MARRIED"
F63	DVRCDEV	"EVER BEEN DIVORCED"
F64	CHLDHM	"CHILDREN LIVING AT HOME OR NOT"
F65	CHLDHHE	"EVER HAD CHILDREN LIVING IN HOUSEHOLD"

Source: Appendix, ESS Data Documentation Report, available form *http://ess.nsd.uib.no/:* ESS Round 1 – Survey documentation.

HARMONISATION OF SURVEY DATA IN THE INTERNATIONAL SOCIAL SURVEY PROGRAMME (ISSP)

*EVI SCHOLZ**

The following article deals with harmonisation in the International Social Survey Programme (ISSP). We start with an introduction into the ISSP and its general concept of harmonisation, then go into harmonisation of the ISSP background variables and give some insights in presenting one major background variable and its construction. The article ends with some brief notes on the future development and prospects for the ISSP background variables.

1 The International Social Survey Programme (ISSP)

The ISSP is a collaborative programme of cross-cultural social science survey research with currently 39 member countries (2004). Each individual ISSP member is responsible for funding and fielding its national ISSP study. The ISSP run an annual module on a topic important for social science research which changes from year to year and is replicated regularly (Scholz, Harkness & Klein, 2003). The ISSP questionnaire consists of two parts: a fixed set of substantive questions with 60 items and since ISSP 2002[1] a second part of obligatory socio-demographic (ISSP standard background) variables (Braun & Uher, 2003).

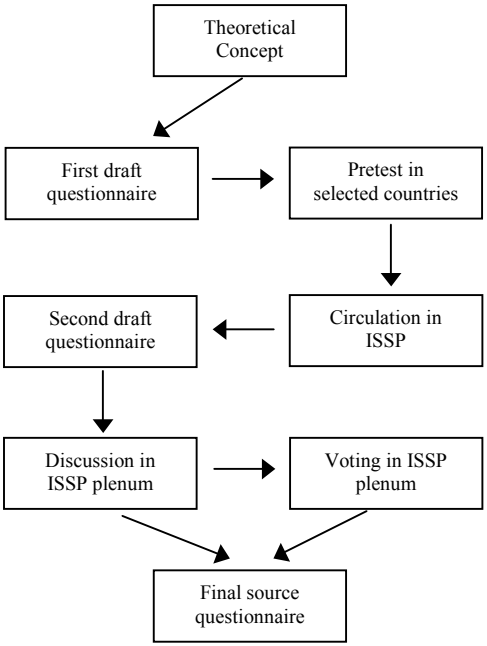
For the substantive part the ISSP applies the concept of input harmonisation: questions are developed, discussed and pretested before they are finally fielded in the ISSP-countries. Questions, answer categories and question order are fixed. Figure 1 illustrates the steps of development of the ISSP's substantive questions. The questionnaire development starts

* I am grateful for the comments from Natalia Garcia-Pardo from CIS, Spain and Kirstine Kolsrud from NSD, Norway regarding the Spanish and Norwegian data. All mistakes, of course, are on my liability.

1 The decision for this fixed set of obligatory background variables was taken in 2001, and came into force with the next following ISSP module.

with theoretical considerations on the topics which should be covered. A drafting group with several ISSP country members' delegates develop a first draft questionnaire which is usually pretested. The draft is circulated together with the pretest results in the whole ISSP for comments and suggestions. A second draft is then developed by the drafting group, discussed in detail at the plenary meeting (the annual general ISSP assembly) and then voted question-by-question using majority rule. The final ISSP source questionnaire results from the discussion and decision at the annual meeting. This questionnaire is then translated and fielded by the individual ISSP member countries.

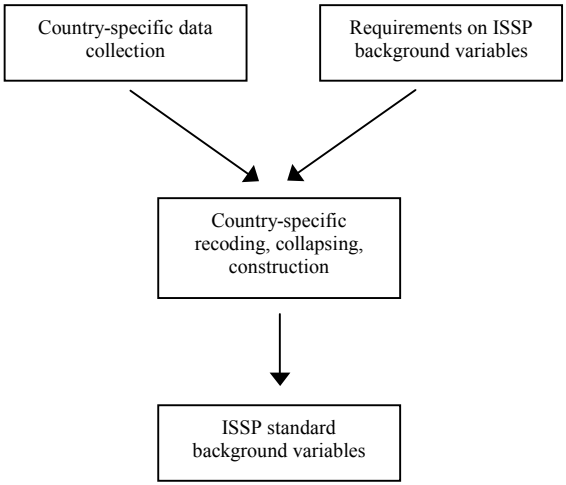
Figure 1 Input Harmonisation in the ISSP: Development of the ISSP Questionnaire with Substantive Questions



For the ISSP standard background variables a mixed mode approach of harmonisation is followed (see Figure 2): The ISSP has currently 34 background variables[2] (see Appendix 1).

2 33 variables are obligatory.

Figure 2 **Mixed Mode Harmonisation in the ISSP: ISSP Standard Background Variables**



These background variables have to be enclosed in all ISSP member countries' questionnaires. All background variables have fixed pre-defined categories which should be covered. But, on the other hand, there are no fixed rules how the individual ISSP members have to ask their background variables, and there is no fixed question wording nor question order. So the ISSP background variables are the result of a country-specific harmonisation process where the outcoming variables are fixed in advance, especially by common categories valid for all ISSP member countries. A paper, mainly produced for internal ISSP use but publicly available[3] lists the ISSP standard background variables and gives some briefing how the ISSP standard background variables should be constructed.

These background variables include sex, age, marital status, religious denomination, a set of variables on the work status of the respondent and a few political background variables.

General problems for the quality of the ISSP background variables come from country-specific restrictions on asking questions, some of them result especially from the particularity that in several countries the ISSP is run together with larger established surveys which cannot change their time-series easily. In these countries ISSP background variables are

3 http://www.za.uni-koeln.de/data/en/issp/codebooks/bv2001.pdf

taken from the larger surveys' background variables or are based on them (Braun & Uher, 2003: 36). Another drawback is that the documentation on background variables is not always as complete as desired. Both problems mostly come from the fact that for the ISSP there are no central funds, particularly not for data preparation or documentation. In the following we will focus on one of the major background variables and their construction in Germany[4], Norway[5], and Spain[6]. The presented variable – highest educational degree (DEGREE) – is taken from ISSP 2002: For the highest educational degree a scheme was developed to facilitate harmonisation of national educational degree (asked country-specifically) into the comparable highest level of educational achievement.

We will start with some general information on the harmonised variable DEGREE, its categories, the briefing given (see Table 1; taken from the guidelines on background variables[7]) and then go into the details of construction.

Education, it goes without saying, can be measured in different ways, but as a comparable measure, not all of them are quite satisfactory: Years of schooling, e.g., produces problems of comparability since the number of school years to achieve a certain qualification is different across countries; additional problems occur in comparing education if some countries allow for repetition of classes, and others don't follow that concept. The ISSP therefore decided to include a variable on education using the country-specific and detailed categories to get the highest educational level and then recode it into a common ISSP educational classification. The final categories of the harmonised variable run from no formal qualification over the lowest formal qualification attainable to a completed university degree. Guidelines on the measurement of education in international projects and details on educational systems and the country-specific classification of education are collected for the ISSP, and are freely available for interested researchers[8].

4 ISSP 2002 in Germany was fielded with ALLBUS 2002: ISSP substantive questions were included in a self completion questionnaire which should be filled in by the respondents with interviewer attendance; the socio-demographic variables were mostly taken or constructed from ALLBUS which was done as CAPI. In the following tables for Germany, the data of the two German subsamples (eastern and western) are collapsed and weighted to adjust the relative size of the two subsamples to an overall representative German sample.

5 ISSP 2002 in Norway was part of a larger mail survey with ISSP at start.

6 Stand alone, face to face.

7 See http://www.gesis.org/en/data_service/issp/rules/background_vars.htm

8 Attachment to the codebook of ISSP 1995: http://www.za.uni-koeln.de/data/en/issp/codebooks/s2880app.pdf.

Table 1 ISSP Highest Educational Degree

Harmonised background variable	DEGREE
Label of harmonised background variable	R: Education II: categories
Briefing on construction	Education II – highest education level / degree- The DEGREE of education should be ASKED country-specific, NOT using the ISSP categories in the questions to the respondents. These ISSP categories should be computed from the country-specific degrees (nat_DEGREE)
Categories	0. No formal qualification 1. Lowest formal qualification attainable 2. Qualifications which are above the lowest qualification, but below the usual entry requirement for universities (intermediary secondary completed) 3. (Usual) Entry requirement for universities (higher secondary completed: the German Abitur, the French Bac, English A-level, etc.) 4. Qualifications which are above the higher secondary level, but below a full university degree 5. University degree completed 9. No answer

Table 2 presents results for the harmonised variable for Germany, Norway and Spain. In the left column the harmonised ISSP categories from Table 1 are repeated, on the right hand side the corresponding frequency distributions for the selected countries are presented. This summary table already shows a possible problem with the lowest category (without formal qualification) since for Norway there are no respondents in that category. The explanation for that Norwegian peculiarity will be given when the variable is deconstructed into their national sources.

Table 2 Highest Educational Degree in ISSP 2002

DEGREE	**Germany in % (n=1367)**	**Norway in % (n=1475)**	**Spain in % (n=2471)**
No formal qualification	3.9	–	12.3
Lowest formal qualification attainable	35.9	10.1	25.0
Qualifications which are above the lowest qualification, but below the usual entry requirement for universities (intermediary secondary completed)	32.1	17.0	26.9
(Usual) Entry requirement for universities (higher secondary completed: the German Abitur, the French Bac, English A-level, etc.)	9.3	34.8	20.1
Qualifications which are above the higher secondary level, but below a full university degree	6.6	11.5	7.5
University degree completed	12.0	26.1	7.9
NA, DK	0.3	0.5	0.3

Source: ISSP 2002, own calculation.

In Germany, the school system is a hierarchical one, where pupils are selected at a quite young age into different kinds of secondary schools (Anweiler, 1996; KMK, 2005). The German General Social Survey (ALLBUS) asks two questions on education, one on the highest general educational attainment, and another one on vocational education and the degree achieved (for full question wording, see Appendix 2). The use of the latter for the construction of DEGREE is necessary since university degree is not asked as school degree but as terminating vocational training. These two variables are combined to one German educational degree variable (also included as obligatory variable D_DEGR in the international ISSP data set): In a first step the answer categories of the question on general educational attainment are collapsed. In a second step respondents reporting a polytechnic or a university degree are coded into two extra categories. Information about other vocational training is not taken into account. Table 3 gives the construction scheme for educational degree in Germany.

Table 3 From German Degree to ISSP Degree (ISSP 2002)

National Degree (D_DEGR)				Recode	ISSP DEGREE	
Educational level	**German Code**	**German Categories**	**English Translation**		**ISSP Code**	**ISSP Categories**
General	1	Noch Schüler	Still at school	→	0	No formal qualification
General	2	Schule beendet ohne Abschluss	Left school without qualification	→		
General	3	Volks-/ Hauptschulabschluss, Polytechnische Oberschule (8./9. Klasse)	Lowest formal qualification of German tripartite secondary school system, after 8 or 9 years of schooling	→	1	Lowest formal qualification
General	4	Mittlere Reife, Realschulabschluss, Polytechnische Oberschule (10. Klasse)	Intermediary secondary qualification, after 10 years of schooling	→	2	Above lowest qualification
General	5	Fachhochschulreife	Certificate fulfilling entrance requirements to study at a polytechnic college	→		
General	6	Abitur bzw. Erweiterte Oberschule (12. Klasse) (Hochschulreife)	Higher qualification, entitling holders to study at a university	→	3	Higher secondary completed
Vocational	7	Fachhochschulabschluss (auch Abschluss einer Ingenieurschule)	Polytechnic degree (or engineering college degree)	→	4	Above higher secondary level but below full university degree
Vocational	8	Hochschulabschluss	University degree	→	5	University degree completed
General	9	anderer Schulabschluss	Other general educational qualification	→	9	NA

Table adapted from: http://www.za.uni-koeln.de/data/en/issp/codebooks/s2880app.pdf

Table 4 shows the corresponding frequency distribution of the harmonised variable DEGREE based on the German combined educational degree variable. Recoding the German highest educational attainment into ISSP educational level fits quite well. The only problem is how to deal with persons where school is not finished at the time of the interview (highlighted in the table). Since respondents are aged 18 and above, usually persons have finished their compulsory education (starting from an age of 6 and ending after 9 or ten classes). So, the respondents "still at school" [9] might include a few cases who could probably be coded into one of the higher categories, e.g., those having passed the 12th class of the German Gymnasium and therefore being equivalent to having a certificate to study at a polytechnic college. But given the low number of persons, the mistake produced here might be negligible. So, the harmonized ISSP variable on highest educational degree for Germany might slightly overrepresent those in the "no formal qualification"-category.

Table 4 Highest Educational Degree in Germany and German ISSP; 2002

German Educational Level	ALLBUS in %	ISSP Category on Highest Educational Degree	ISSP in %
Still at school	1.1	No formal qualification	3.9
Finished school without school leaving certificate	2.8		
Lowest formal qualification of Germany's tripartite secondary school system, after 8 or 9 years of schooling	35.9	Lowest formal qualification attainable	35.9
Intermediary secondary qualification, after 10 years of schooling	27.7	Qualifications which are above the lowest qualification, but below the usual entry requirement for universities (intermediary secondary completed)	32.1
Certificate fulfilling entrance requirements to study at a polytechnic college	4.4		
Higher qualification, entitling holders to study at a university	9.3	(Usual) Entry requirement for universities (higher secondary completed: the German Abitur, the French Bac, English A-level, etc.)	9.3
Polytechnic degree (or engineering college degree)	6.6	Qualifications which are above the higher secondary level, but below a full university degree	6.6
University degree	12.0	University degree completed	12.0
NA, DK	0.3	NA, DK	0.3

Source: ISSP 2002, n=1367, own calculation.

9 Still at school: n=15 or 1.1 %; thereof 4 are not born in Germany and immigrated between 1990 and 1995; 2 of the 15 pupils reported that they are less than half time employed. No further information on type of school.

During the last decades, the Norwegian school system has experienced profound reforms, new educational institutions have been founded and others previously existing have been changed. Vocational and academic educational programmes have been integrated into one system. The number of years in compulsory education has increased from 7 years to 10 years nowadays. Access to higher education is usually obtained by a three-year course of upper secondary education, either general or vocational. Higher education in Norway is provided at colleges or universities, some higher educational programmes last one to three years but the degree programmes take three to seven years, in general. In the pre-reform system, persons with university degree started their educational career with a compulsory primary school, then attended a not compulsory lower secondary school (*realskole*), followed by a *gymnas* at upper secondary level and university studies. Persons without academic interests after completed primary education were prepared for vocational training in the so-called continuation school (*framhaldsskole*) followed by job training (Aakvik, Salvanes & Vaage, 2003; Eurydice, 2003a; ISSP, 1995; Statistics Norway, 2003).

The highest educational level is constructed from three separate questions; the first one on general education, a second one on vocational and a third one on higher education (for full question wording, see Appendix 2) covering different educational backgrounds of different ages groups. These three variables are combined into one Norwegian educational degree variable (named N_DEGR and included in the international data set). Table 5 gives the construction scheme.

Table 6 shows the corresponding frequency distribution of the harmonised variable DEGREE based on the Norwegian combined educational degree variable and demonstrates where the empty category of the harmonised variable comes from: The, at first glance, missing respondents with no formal qualification result from the Norwegian educational system where pupils automatically move from one class to the next during their primary education[10]. So dropouts are not expected to occur or if, then only in small numbers.

The Spanish educational system has experienced basic changes in the post-Franco years. Until 1970, school education in Spain was influenced by the *Ley Moyano* a more than one hundred years old legal regulating where, for example, elementary education was offered for children between 6 and 9 years and charged with fees (Goetze, 1996: 213). With the educational reform process starting in 1970, education becomes as a public issue, where general and vocational education is obligatory and without fees.

10 During the first and second stage of primary education, there is no formal assessment. In the third stage of primary education, marks for each subject based on the teachers' assessment are awarded twice a year. But promotion to the next grade is automatic (Eurydice, 2003a).

Table 5 From Norwegian Degree to ISSP Degree (ISSP 2002)

National Degree (questionnaire categories and N_DEGR)						Recode	ISSP DEGREE	
Educational Level	Norwegian Questionnaire Categories	English Translation	Recode	N_DEGREE Categories	N_DEGREE Code		ISSP Code	ISSP Categories
General	-	-			-		0	No formal qualification
General	7-årig folkesskole eller kortere	Elementary education completed (compulsory, 7 years or less)		Primary completed	1		1	Lowest formal qualification
General	9-eller 10-årig grunnskole	Basic compulsory education completed (9 or 10 years)						
Vocational	Grunnutdanning ved yrkesskole, videregående skole	Secondary vocational incomplete (education at vocational or secondary school)		Secondary vocational incomplete	2		2	Above lowest qualification
General	Framhaldsskole eller fortsellelsesskole	Secondary academic incomplete (secondary school)		Secondary academic incomplete	3			
General	Realskole eller middelskole	Secondary academic incomplete (intermediary secondary school)						
General	Folkehøgskole	Secondary academic incomplete (high school, not leading to an academic degree)						
Vocational	Fagbrev eller fullført fagutdanning ved yrkesskole, handelsskole, yrkesfag ved videregående skole	Secondary vocational completed (certificate of apprenticeship or technical education completed, business school, vocational education at secondary school completed		Secondary vocational completed	4		3	Higher secondary completed
General	Eksamen artium, økonomisk gmynas eller 3-årig videregående skole med generell studiekompetanse	Secondary academic completed (high school degree, economic high school, or 3-years secondary school with general entry requirement for university)		Secondary academic completed	5			
Higher	Offentlig godkjent høyskole eller universitet av mindre enn 1 års varighet	Publicly approved university less than 1 year		University, college < 1 year	6		4	Above higher secondary level but below full university degree
Higher	Offentlig godkjent høyskole eller universitet av 1-2 års normert varighet	Publicly approved university 1-2 years		University, college 1-2 years	7			
Higher	Offentlig godkjent høyskole eller universitet av 3-4 års normert varighet	Publicly approved university 3-4 years		University, college 3-4 years	8		5	University degree completed
Higher	Offentlig godkjent høyskole eller universitet av 5 års normert varighet eller mer	Publicly approved university 5 years or more		University, college>=5 years	9			
	Not included in questionnaire			NA, DK	99		9	NA

Table adapted from: http://www.za.uni-koeln.de/data/en/issp/codebooks/s2880app.pdf

Higher education is offered at universities, technical colleges (length of studies 5-6 years) or at higher polytechnic schools comparable to the German *Fachhochschulen* for the education of primary school teachers, translators, opticians, etc.[11] (length of studies 3 years). University education is organised in three cycles, a first basic one of three years, a second one of two additional years with degree *licenciatura* and doctoral studies as the third cycle (Eurydice, 2003b; Goetze, 1996).

Table 6 Highest Educational Degree in Norway and Norwegian ISSP; 2002

Norwegian Educational Level	Norway in %	ISSP Highest Educational Degree	ISSP in %
–	–	No formal qualification	–
Primary completed	10.1	Lowest formal qualification attainable	10.1
Secondary vocational incomplete Secondary academic incomplete	6.8 10.2	Qualifications which are above the lowest qualification, but below the usual entry requirement for universities (intermediary secondary completed)	17.0
Secondary vocational completed Secondary academic completed	17.2 17.7	(Usual) Entry requirement for universities (higher secondary completed: the German Abitur, the French Bac, English A-level, etc.)	34.8
University college < 1 year University college 1-2 years	2.8 8.7	Qualifications which are above the higher secondary level, but below a full university degree	11.5
University college 3-4 years University college >=5 years	16.9 9.2	University degree completed	26.1
NA, DK	0.5	NA, DK	0.5

Source: ISSP 2002, n=1475, own calculation.

Since Spaniards born during the time of the *Ley Moyano* and those born in the post-Franco time were educated in different educational systems, Spanish researchers have to deal with those different educational histories. Hence, in Spain, education is also asked very detailed. The first question is on school attendance in general (whether the respondent is illiterate, didn't attend school but is able to red or write or attended school), the second one asks the highest level, specifying the grade and the name of the studies fin-

11 Architects and engineers educated for three years in polytechnic schools achieve a different degree from those educated in technical colleges.

ished. The third question is on the number of years of official schooling (for full question wording, see Appendix 2). The information achieved is then coded by the Spanish research team. The outcome is stored in a Spanish educational degree variable (named E_DEGR, for full wording, Appendix 2) and serves as the base for the construction of DEGREE presented in Tables 7 and 8.

As the frequency distribution shows, the number of people without completed primary education is quite high resulting from the influence of the pre 1970 school education in Spain which effects the educational level of many Spaniards (Goetze 1996: 213).

Table 7 From Spanish Degree to ISSP Degree (ISSP 2002)

National Degree (E_DEGR)			Recode	ISSP DEGREE	
Spanish Code	**Spanish Categories**	**English Translation**		**ISSP Code**	**ISSP Categories**
1	Ninguna[12]	None	→	0	No formal qualification
2	Educación básica incompleta	Incomplete primary school	→		
3	Educación básica completa	Primary school completed	→	1	Lowest formal qualification
4	Educación secundaria incompleta	Incomplete secondary school	→	2	Above lowest qualification
5	Formación profesional I y enseñanzas técnico-profesionales equivalentes	Basic level of vocational education	→		
6	Educación secundaria completa	Secondary school	→	3	Higher secondary completed
7	Formación profesional II y equivalentes	Secondary level of vocational education	→		
8	Arquitecto e ingeniero técnico	Technical architecture & engineering	→	4	Above higher secondary level but below full university degree
9	Estudios universitarios incompletos	3 years university	→		
10	Arquitecto e Ingeniero Superior	Architecture and engineering academic	→	5	University degree completed
11	Carrera universitaria completa, Licenciatura	College or university first degree	→		
12	Estudios de Postgrado o especialización	PhD, other official degree	→		
99	N.C.	NA, DK		9	NA

12 Either illiterate or no school attendance, but able to read or write.

Table 8 Highest Educational Degree in Spain and Spanish ISSP; 2002

Spanish Educational Level	Spain in %	ISSP Highest Educational Degree	ISSP in %
None	6.8	No formal qualification[13]	12.3
Incomplete primary school	5.5		
Primary school completed	25.0	Lowest formal qualification attainable	25.0
Incomplete secondary school	22.1	Qualifications which are above the lowest qualification, but below the usual entry requirement for universities (intermediary secondary completed)	26.9
Basic level of vocational education	4.7		
Secondary school	13.3	(Usual) Entry requirement for universities (higher secondary completed: the German Abitur, the French Bac, English A-level, etc.)	20.1
Secondary level of vocational education	6.8		
Technical architecture & engineering	1.5	Qualifications which are above the higher secondary level, but below a full university degree	7.5
3 years university	6.1		
Architecture & engineering academic	1.0	University degree completed	7.9
College or university first degree	5.4		
PhD, other official degree	1.5		
NA, DK	0.3	NA, DK	0.3

Source: ISSP 2002, n=2471, own calculation.

To conclude, to measure formal school education, DEGREE is one of the well-constructed ISSP background variables and does not produce remarkable problems. Special efforts have been made to develop guidelines, to design the variable and to document the bridges from the national educational variables to the harmonised ISSP variable.

2 Conclusion and Prospects

Harmonisation of variables in cross-national surveys is always a demanding enterprise and a compromise. In a sense the harmonisation of background variables is also a puzzle which fits more or less well. The ISSP asks background variables which are well constructed but also others which might produce problems in terms of comparability. The experiences from ISSP 2002, the first ISSP module applying the ISSP standard background variables, have resulted in an ISSP internal review process to improve the quality of the ISSP background variables. In a first step, all ISSP members were asked to send comments on the ISSP demographic variables to its demographic methods work group (DMG) to get insights where the ISSP member countries themselves feel there is potential for modifications, changes, replacements or even cuts of variables. All comments were collected and put together in a report which includes the DMG's recommendations based

13 About 78 % of respondents without formal qualification are aged 60 and above.

on the ISSP's member reaction. The report was circulated at the end of 2004 and needs to be discussed during the next months. A next step of a probably long-lasting discussion and decision-making process might be a complete documentation of the ISSP background variables and their construction in each individual member country. Such a detailed documentation is a necessary pre-condition for quality assessment and, in a more distant future, for a modification of the obligatory ISSP standards on its demographic variables. For established national surveys, any changes of institutionalised procedures, questions and variables are difficult where a time-series would be interrupted. The difficulties for change, of course, increase in international survey programmes such as the ISSP, where several countries run the ISSP together with its General Social Survey (such as Australia, Great Britain, Germany; Poland or the USA) and with more than half of all countries delivering ISSP data fielded together with a larger survey (Klein & Harkness, 2004).

References

Aakvik, A., K. G. Salvanes, and K. Vaage. 2003. *Measuring Heterogeneity in the Returns to Education in Norway Using Educational Reforms,* Norwegian School of Economics and Business Administration, Department of Economics, Discussion Paper 26/2002. http://www.nhh.no/sam/res&publ/2002/dp26.pdf (23.3.2005).

Anweiler, O. 1996. „Deutschland." S. 13-30 in *Bildungssysteme in Europa. Entwicklung und Struktur des Bildungswesens in zehn Ländern: Deutschland, England, Frankreich, Italien, Niederlande, Polen, Rußland, Schweden, Spanien, Türkei,* herausgegeben von O. Anweiler et al. 4. revised edition. Weinheim and Basel: Beltz Verlag.

Braun, M., and P. Ph. Mohler. 2003. „Background Variables." Pp. 101-115 in *Cross-cultural Survey Methods,* edited by J. A. Harkness, F. J. R. De Vijver and P. Ph. Mohler. Hoboken. N. J.: John Wiley & Sons.

Braun, M., and R. Uher. 2003. "The ISSP and its Approach to Background Variables." Pp. 33-47 in *Advances in Cross-National Comparison,* edited by Hoffmeyer-Zlotnik, J. H.P., and C. Wolf. New York: Kluwer 2003.

Hoffmeyer-Zlotnik, J. H.P., and C. Wolf (Eds.). 2003. *Advances in Cross-national Comparison.* New York: Kluwer Academic, Plenum Publishers.

Eurydice. 2003a. *Structures of Education, Vocational Training, and Adult Education, Systems in Europe.* Norway 2003, information provided by the Royal Norwegian Ministry of Education and Research, Eurydice Unit. http://www.eurydice.org (28.01.2005).

Eurydice. 2003b. *Structures of Education, Vocational Training, and Adult Education, Systems in Europe*. Spain 2003, information provided by Eurydice Unit, Ministry of Education, Culture and Sport CIDE-Centro de Investigación y Documentación. http://www.eurydice.org (28.01.2005).

Goetze, D. 1996. „Spanien“. S. 213-230 in *Bildungssysteme in Europa. Entwicklung und Struktur des Bildungswesens in zehn Ländern: Deutschland, England, Frankreich, Italien, Niederlande, Polen, Rußland, Schweden, Spanien, Türkei,* herausgegeben von O. Anweiler et al. 4. revised edition. Weinheim and Basel: Beltz Verlag.

Harkness, J. and M. Blohm. 2004. “German General Social Survey 2002, English Translation of the ALLBUS Questionnaire.” *ZUMA-Methodenbericht 2004/3*. http://www.gesis.org/Dauerbeobachtung/Allbus/documents/pdfs/GGSSquest2002.pdf (16.12.2004).

ISSP. 1995. *Codebook*. Appendix. http://www.za.uni-koeln.de/data/en/issp/codebooks/s2880app.pdf (17.3.2005).

Klein, S. and J. Harkness. 2004. “ISSP Study Monitoring 2002. Report to the ISSP General Assembly on monitoring work undertaken for the ISSP by ZUMA, Germany.” *ZUMA Methodenbericht 2004/10.*

KMK 2005. *Das Bildungswesen in der Bundesrepublik Deutschland 2002. Darstellung der Kompetenzen, Strukturen und bildungspolitischen Entwicklungen für den Informationsaustausch in Europa,* herausgegeben vom Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland, Bonn. http://www.kmk.org/dossier/dossier_2002/dossier_2002_dt_ebook.pdf (28.01.2005).

Norway. 2005. *Norway, the official site in the United States*. http://www.norway.org (17.3.2005).

Scholz, E., J. Harkness, and S. Klein. 2003. “ISSP 2002 Germany, Family and Changing Gender Roles III, ZUMA Report on the German Study.” *ZUMA Methodenbericht 2003/15.*

Statistics Norway. 2003. *Norwegian Standard Classification of Education, Revised 2000.* http://www.ssb.no/english/subjects/04/90/nos_c751_en/nos_c751_en.pdf (28.01.2005).

Appendix 1 ISSP Standard Background Variables

SEX	Sex of respondent
AGE	Age of respondent
MARITAL	Marital status of respondent
COHAB	Do you live together with a partner?
EDUCYRS	Education I – years (of full time) schooling
DEGREE	Education II – highest education level / degree
nat_DEGR	Education II – country specific categories
WRKST	Current employment status, main source of living
WRKHRS	Working hours – number of hours (usually) worked weekly *(including regular overtime work)*
ISCO88	Occupation – four digit 1988 ISCO / ILO occupation code *(if R not working ask most recent job)*
WRKSUP	Supervises others at work – Do you supervise the work of any other people? *(if R not working ask most recent job)*
WRKTYPE	R: Working for private sector, public sector or self-employed *(if R not working ask most recent job)*
NEMPLOY	(If self employed) Do you have any employees? If so, how many? *(if R not working ask most recent job)*
UNION	Is respondent member of a trade union *(or was in the past)?*
SPWRKST	Spouse / partner: current employment status, main source of living
SPISCO88	Spouse / partner: occupation, four digit 1988 ISCO / ILO occupation code *(if R not working ask most recent job)*
SPWRKTYP	S-P: Working for private sector, public sector or self-employed *(if R not working ask most recent job)*
INCOME	Family income
RINCOME	Respondent's earnings
HOMPOP	How many persons in household
HHCYCLE	Household composition: adults and children (less than 18 years)
PARTY_LR	Party affiliation – left-right *(Not asked but coded in a 5-point left - right scheme by data producers)*
nat_PRTY	Party affiliation – Country-specific
VOTE_LE	Vote last general election: Yes/no
ATTEND	Attendance of religious services
RELIG	Religious denomination *(asked country-specific but re-coded to standard)*
RELIGGRP	Religious main-group derived from RELIG
TOPBOT	10 point top-bottom scale
nat_REG	Region – Country-specific
nat_SIZE	Size of community – Country-specific
URBRURAL	Type of community: urban / rural – self-assessed
WEIGHT	Weighting factor
MODE	Administrative mode of data-collection
nat_ETHN (optional)	*Ethnic identity / family origin (Country specific)*

Appendix 2 Original Question Wording and their English Translation
(Instructions in italics)

DEGREE: **German Question Wording, Original**	**DEGREE:** **German Question Wording, English Translation**
S3 *INT: Liste S 3 vorlegen!* Als nächstes kommen jetzt Fragen zu Ihrer Ausbildung, Ihrem Beruf und Ihrer Familie. Beginnen wir mit Ihrer Ausbildung: Welchen allgemeinbildenden Schulabschluss haben Sie? *INT: Nur höchsten Schulabschluss angeben lassen!* 1: A Noch Schüler 2: B Schule beendet ohne Abschluss 3: C Volks-/ Hauptschulabschluss bzw. Polytechnische Oberschule mit Abschluss 8. oder 9. Klasse 4: D Mittlere Reife, Realschulabschluss bzw. Polytechnische Oberschule mit Abschluss 10. Klasse 5: E Fachhochschulreife (Abschluss einer Fachoberschule etc.) 6: F Abitur bzw. Erweiterte Oberschule mit Abschluss 12. Klasse (Hochschulreife) 7: G Anderen Schulabschluss, und zwar _____ 9 Keine Angabe	S3 *INT: Please display card S3!* Next we come to questions about your education, job and family. Let's begin with your education. What general school leaving certificate do you have? *INT: Please request highest school leaving certificate only* 1 A Still at school 2 B Finished school without school leaving certificate 3 C Lowest formal qualification of Germany's tripartite secondary school system, after 8 or 9 years of schooling 4 D Intermediary secondary qualification, after 10 years of schooling 5 E Certificate fulfilling entrance requirements to study at a polytechnic college 6 F Higher qualification, entitling holders to study at a university 7 G Other school leaving certificate, _____ 9 No answer
S4_j / S4_k (nicht für Schüler) *INT: Liste S 4 vorlegen!* Welchen beruflichen Ausbildungsabschluss haben Sie? Was von dieser Liste trifft auf Sie zu? Nennen Sie mir bitte die entsprechenden Kennbuchstaben. *INT: Mehrfachnennungen möglich, außer wenn „M" genannt.* 01 A Beruflich-betriebliche Anlernzeit mit Abschlusszeugnis, aber keine Lehre 02 B Teilfacharbeiterabschluss 03 C Abgeschlossene gewerbliche oder landwirtschaftliche Lehre 04 D Abgeschlossene kaufmännische Lehre 05 E Berufliches Praktikum, Volontariat 06 F Berufsfachschulabschluss 07 G Fachschulabschluss 08 H Meister-, Techniker- oder gleichwertiger Fachschulabschluss 09: J Fachhochschulabschluss (auch Abschluss einer Ingenieurschule)* 10: K Hochschulabschluss* 11: L Anderen beruflichen Ausbildungsabschluss, und zwar… 12: M Keinen beruflichen Ausbildungsabschluss	S4_j / S4_k (if not still at school) *INT: Please display list S4!* What vocational or professional training do you have? Which of the categories on the card apply to you? Please name the appropriate letter(s). *INT: Multiple responses possible, unless "M" selected.* 01 A On-the-job vocational training with final certificate, but not within traineeship or apprenticeship scheme 02 B Compact vocational training course 03 C Completed trades/crafts or agricultural traineeship 04 D Completed commercial traineeship 05 E Work placement/internship 06 F Technical or vocational college certificate 07 G Specialized vocational college certificate 08 H Master(craftsman), technician or equivalent college certificate 09 J Polytechnic degree (or engineering college degree)* 10 K University degree* 11 L Other vocational training certificate, please enter 12. M No completed vocational training

* Items used for the construction of DEGREE in Germany
Translations by Sylvia Hönig, taken from Harkness & Blohm (2004)

Appendix 2 (continued)

DEGREE: **Norwegian Question Wording, Original**	**DEGREE:** **Norwegian Question Wording, English Translation**
70. Hvilken allmennutdanning har du fullført? *Kun ett kryss mulig* 7-årig folkeskole eller kortere Framhaldsskole eller fortsellelsesskole 9-eller 10-årig grunnskole Realskole eller middelskole Folkehøgskole[14] Eksamen artium, økonomisk gmynas eller 3-årig videregående skole med generell studiekompetanse	What kind of general education have you completed? *Please tick* Elementary education completed (compulsory, 7 years or less) Secondary academic incomplete (secondary school) Basic compulsory education completed (9 or 10 years) Secondary academic incomplete (intermediary secondary school) Secondary academic incomplete (high school, not leading to an academic degree) Secondary academic completed (high school degree, economic high school, or 3-years secondary school with general entry requirement for university)
71. Hvilken yrkesutdanning har du fullført? *Kun ett kryss mulig* Ingen Grunnutdanning ved yrkesskole, videregående skole Fagbrev eller fullført fagutdanning ved yrkesskole, handelsskole, yrkesfag ved videregående skole	What is the highest vocational education that have you completed? *Please tick* None Secondary vocational incomplete (education at vocational or secondary school) Secondary vocational completed (certificate of apprenticeship or technical education completed, buisness school, vocational education at secondary school completed
72. Hvilken høyere utdanning har du fullført? *Kun ett kryss mulig* Ingen Offentlig godkjent høyskole eller universitet av mindre enn 1 års varighet Offentlig godkjent høyskole eller universitet av 1-2 års normert varighet Offentlig godkjent høyskole eller universitet av 3-4 års normert varighet Offentlig godkjent høyskole eller universitet av 5 års normert varighet eller mer	What is the highest education at university/college that have you completed? *Please tick* None Publicly approved university less than 1 year Publicly approved university 1-2 years Publicly approved university 3-4 years Publicly approved university 5 years or more

14 Folkehøgskole (Folk high school) are mostly run by private organisations, offer an overall educational programme to develop individual, social and academic skills. Folk high schools do not conduct formal exams. They are mostly attended by young adults having completed their upper secondary education, especially by those looking for a transitional year (source: Norway (2005), Norway, the official site in the United States, http://www.norway.org).

Appendix 2 (concluded)

DEGREE: **Spanish Question Wording, Original**	**DEGREE:** **Spanish Question Wording, English Translation**
P26 ¿Ha ido Ud. a la escuela o cursado algún tipo de studios? *(ENTREVISTDOR: EN CASO NEGATIVO, PREGUNTAR SO SABE LEER Y ESCRIBIR).* No. Es analfabeto No, pero sabe leer y escribir Si, ha ido a la escuela N.C.	Have you attended school or taken any kind of courses? *(Interviewer: if negative, ask whether R can read and write)* No, R is illiterate No, but able to read and write Yes, R attended school NA
P26a. ¿Cualés son los estudios de más alto nivel oficial que Ud. ha cursado (con independencia de que los haya terminado o no)? Por favor, especifique lo más posible, diciéndome el curso en que estaba cuando los terminó (o los interrumpió), y también el nombre que tenían entonces esos estudios. (ej: 3 años de Estudios Primarios, Primaria, 5° de Bachilerato, Maestría Industrial, Preuniversitario, 4° de EGB, Licenciatura, Doctorado, FP1, etc.). *(ENTREVISTDOR: SI AÚN ESTÁ ESTUDIANDO, ANOTAR EL ÚLTIMO CURSO QUE HAYA COMPLETADO. SI NO HA <u>COMPLETADO</u> LA PRIMARIA, ANOTAR N° DE AÑOS QUE ASISTÍO A LA ESCUELA).* CURSO… NOMBRE (de los estudios)… NIVEL (Codificar según T. ESTDUIOS)… P26b. ¿Cuántos años estuvo Ud. esolarizado o cursando estudios oficiales a los que dedicara la jornada completa? …años Todavía está estudiando (colegio/instituto) Todavía está estudiando (universidad) No recuerda N.C.	What is the highest level of schooling you have ever taken (independent of whether you finished theses studies or not)? Please specify as much as possible, telling me the grade you were in when you finished (or interrupted) and the name of the schooling level . (i.e., 3 years of primary school, Primary School, 5th grade of high school, College Degree, etc) *(Interviewer: if still studying, write down the last completed course. If primary not completed, write down number of years in school)* Grade (first, second, third course etc) Name (of studies) Level (Code according …) How many years of full-time official schooling did you attend? …years Still studying (high school) Still studying (college, university) Don't remember NA

Outcome of detailed questions in Spain

Spanish Educational Level, Original	**Spanish Educational Level, English Translation**
Ninguna	None
Educación básica incompleta	Incomplete primary school
Educación básica completa	Primary school completed
Educación secundaria incompleta	Incomplete secondary school
Formación profesional I y enseñanzas técnico-profesionales equivalentes	Basic level of vocational education
Educación secundaria completa	Secondary school
Formación profesional II y equivalentes	Secondary level of vocational education
Arquitecto e ingeniero técnico	Technical architecture & engineering
Estudios universitarios incompletos	3 years university
Arquitecto e Ingeniero Superior	Architecture and engineering academic
Carrera universitaria completa, Licenciatura	College or university first degree
Doctorate, other official graduate degrees	PhD, other official degree

PART IV: SOCIO-ECONOMIC VARIABLES IN CROSS-NATIONAL PERSPECTIVE

MEASURING INCOME IN COMPARATIVE SOCIAL SURVEY RESEARCH

*UWE WARNER & JÜRGEN H.P. HOFFMEYER-ZLOTNIK**

1 Introduction

Different surveys use different strategies to arrange the questionnaires and different accuracy to construct the income questions. This degree of precision depends on the research interest and the aims of the study.

Market research is interested in categorizing the purchasing power of a household and classifies the households into consumer groups. They focus on classes of income size and therefore they do not give a precise definition of income and they make no distinction between several surveyed population groups. In Germany, they ask for the monthly net income and they give a general instruction in the question wording. The answers are income brackets. In case the interviewed person refuses to answer, the interviewer often takes the freedom to estimate the household's income.

Social research uses income as a socio-economic indicator on social stratification and inequality. From this point of view the knowledge of size classes of the household income is sufficient. But social research defines the various income types and formulates separate questions for different population groups, for example the wording of the income question differs for the self-employed and for employees. In Germany, the monthly net income is surveyed by an open question and/or with income brackets and the amount is given.

Economic and socio-economic research is studying income distribution and the dynamics of changes in the economic situation of the respondent. The research question on how the total income is composed by it components and changes of the income types are of interest.

* This paper is supported by the national science foundation of Luxembourg (FNR) by the contract No. FNR/04/MA6/10, the infrastructures and the "Longitudinal Data Bases and Archive" of CEPS/INSTEAD at Differdange, Luxembourg.

Therefore a precise measurement of income is needed. The several types of income are defined in detail and separated by their sources and types. Specific population groups and/or income recipients are interviewed according their characteristics. For a well-defined time period (e.g. monthly) gross and net income are asked through open questions and all other monetary resources of all persons living in the household, as well as payments to the household per se are asked for. In general the answer is given as a gross and/or net amount.

Subsequent we will compare two different instruments of income measurement.

2 The Surveys

The *"European Community Household Panel" (ECHP)* is a longitudinal study coordinated by Eurostat that follows the same individuals and households over time. The major aims of ECHP are to provide micro-data on household and person level about the income, the monetary well-being and the dynamics of the economic situation in the European community and its member states.

The household questionnaire of the 8th wave covered five income items.

The person questionnaire of the same wave asks for 50 different income objects.

Here we use the user data base version April 2004 available to the academic community. The 8th wave's interviews are carried out in 2001 and refer to the income reference year 2000.

In 15 EU countries 59,852 households with 121,122 members are surveyed.

The *"European Social Survey" (ESS)* is a pan-European cross sectional time series running every two years. In the 2002 survey, 23 countries participated and collected information on people's social attitudes, beliefs in values, social and political behavior.

In each participation country, the survey design of ESS is a random sample with a known inclusion probability of the selected contact person eligible for the interview. Only on household member aged 16 and over is asked; this person also answers the question about the household situation.

For 21 countries 40,856 responses are included into the data-base[1].

1 Data base version date is Feb. 03, 2004

3 The Income Questions

The ESS question wording is: "… if you add up the income from all sources, which letter describes your household's total net income? If you don't know the exact figure, please give an estimate. Use the part of the card that you know best: weekly, monthly or annual income." (ESS 01/08/2002: F30) The interviewer hands over to the respondent a show card with answer categories:

Figure 1 ESS Card "Household Income"

CARD 56

YOUR HOUSEHOLD INCOME

	Approximate WEEKLY	Approximate MONTHLY	Approximate ANNUAL	
J	Less than €40	Less than € 150	Less than €1800	J
R	€40 to under €70	€150 to under €300	€1800 to under €3600	R
C	€70 to under €120	€300 to under €500	€3600 to under €6000	C
M	€120 to under €230	€500 to under €1000	€6000 to under €12000	M
F	€230 to under €350	€1000 to under €1500	€12000 to under €18000	F
S	€350 to under €460	€1500 to under €2000	€18000 to under €24000	S
K	€460 to under €580	€2000 to under €2500	€24000 to under €30000	K
P	€580 to under €690	€2500 to under €3000	€30000 to under €36000	P
D	€690 to under €1150	€3000 to under €5000	€36000 to under €60000	D
H	€1150 to under €1730	€5000 to under €7500	€60000 to under €90000	H
U	€1730 to under €2310	€7500 to under €10000	€90000 to under €120000	U
N	€2310 or more	€10000 or more	€120000 or more	N

Source: ESS 01/08/2002: Card56

Additional explanations are given to the interviewer at the end of the "project instructions": At the income question "you should obtain the *total net income* of the household from all sources, that is, *after tax*. Income includes not only earnings but state benefits, occupational and other pensions, unearned income such as interest from savings, rent, etc.

We want figures *after* deductions of income tax, national insurance, contributory pension payments and so on. The questions refer to *current level* of income or earnings or, if that is convenient, to the nearest *tax* or other period for which the respondent is able to answer. The respondent is given a show card that enables them to choose between their weekly, monthly or annual income, whichever they find easiest. They will then give you the letter that corresponds to the appropriate amount. This system is designed to reassure the respondent about the confidentiality of the information they are giving." (ESS 15/07/2002: 21)

The ECHP measures income by using a sixteen page long section in the person's questionnaire. Every member (fifteen years and over) of an eligible household answers the person questionnaire. The first approach to income is a monthly calendar about the labor force status of the respondent. For the year prior to the year of the interview, month by month the employment situation is collected. (e.g. the eighth wave interviews carried out in 2001 ask about the situation in 2000). The second step forward to the incomes is a sequence on having or not various income sources. After this the respondent is asked to give net and/or gross amounts of his/her income details during the income reference year.

This list summarizes the income types mentioned in the ECHP interviews:

- as an employee
- self-employment
- income and benefits from sources other than work
- pensions
- private transfer
- capital
- reimbursement

One household member, considered as a reference person for the whole household, is also surveyed by a household questionnaire. Five pages of this questionnaire deal with incomes of the household.

4 First Analysis

Table 1 Distribution of Categorized Annual Income by Survey in Selected Countries and for Selected Income Categories (percent)

Survey	Germany	United Kingdom	Italy	Luxembourg
categories 1 to 3 (up to 6,000€)				
ESS	3.7	6.0	8.8	2.3
ECHP8	1.8	2.3	7.0	0.2
category 6 (from 18,000€ to 24,000€)				
ESS	18.4	12.6	16.5	12.0
ECHP8	16.1	12.8	16.8	11.7
category 9 (from 36,000€ to 60,000€)				
ESS	16.0	19.8	10.8	26.1
ECHP8	21.3	26.9	9.9	33.9
category 10 to 12 (from 60,000€ to 120,000€ and more)				
ESS	7.7	15.5	4.5	19.2
ECHP8	4.0	10.7	1.5	23.4

Table 2 Mode and Median of Categorized Annual Income by Survey in Selected Countries

Survey	Germany		United Kingdom		Italy		Luxembourg	
	Mode	Median	Mode	Median	Mode	Median	Mode	Median
ESS	6	7	9	7	4	6	9	8
ECHP8	9	7	9	7	5	5	9	9

Germany

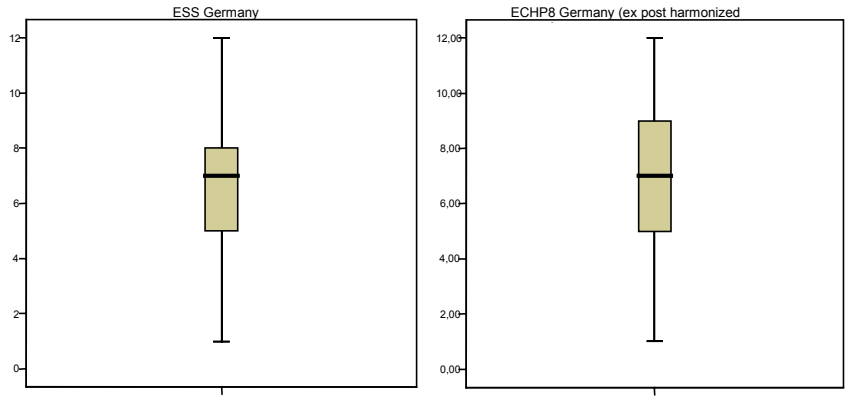
In Germany, 3.7% of the ESS respondents tick the lowest three income categories (up to 6,000€ per year), the ECHP answers of the wave 8 add up to 1.8% of the households having the lowest income categories.

47% of the households surveyed in the ESS have an annul income up to 24,000€, this are 5% points more then households answering the ECHP wave 8 questionnaire.

For the ESS we find the mode at the income range of 18,000 to 24,000€ and the median at the income group of 24,000 to 30,000€, for the ECHP wave 8 the mode as well as the median are in the seventh category where the household has an annual income of 24,000 to 30,000€.

16% of the ESS household have an income of 36,000 to 60,000€, 21% of the ECHP wave 8 households have the same monetary resource.

Figure 2 Box Plot "Household's Total Net Income Germany ESS and ECHP8"



Looking at the upper end of the income categories, the ESS has nearly 8% of the observed households, the 8th wave of ECHP reports 4% of the households having 60,000€ and more annual income. In Germany this group of households at the upper end of the income distribution is small, but comparing both surveys this population is twice as big in ESS then in ECHP wave 8.

In Germany, the respondents of ESS overestimate there total household income at the lower and upper extremes of the income distribution in reference to the ECHP8; at the lower end about 2% of the answering persons and at the upper end about 4% of the interviewed people. In the middle part of the income groups both surveys show the same results.

United Kingdom

In ESS the income categories up to 6,000€ annually are three times often answered as in ECHP wave 8 (ESS = 6% and ECHP8 = 2.2%).

The cumulative frequencies for the categories 1 to 6 (up to 24,000€) differ about 6% between both surveys (ESS = 46% and ECHP8 = 40%).

ESS and ECHP8 have the median at category 7 (24,000 to 30,000€) and the mode at category 9 (36,000 to 60,000€).

19% of the ESS respondents in the United Kingdom have a total annual net household income from 36,000 to 60,000€. The ECHP8 reports nearly 29% of the household in the same category.

At the upper end of the income categories (60,000€ and more) both surveys differ at 50% of the observed cases. In ESS 15% of the surveyed households answer in these categories. In ECHP8 10% of the households are in these income group.

In general, the upper income classes are more frequent in United Kingdom as in Germany.

Respondents, living in households with household income at the bottom or the top end of the income scale, overestimate the total household income; the interviewed persons in the middle categories underestimate their household revenue.

Comparing the cumulative frequency distributions of the two studies, up to the income category 6 (18,000 to 24,000€) the ESS has about 7% more cases in each group as the ECHP8. At the income group 9 (36,000 to 60,000€) the ECHP8 reports more often answers as the ESS.

Table 3 Cumulative Frequencies of Total Net Household Income for United Kingdom

Income Categories				ESS	ECHP wave 8
1:		up to	1,800	0.8	0.5
2:	1,800	–	3,600	2.6	1.0
3:	3,600	–	6,000	6.0	2.3
4:	6,000	–	12,000	22.3	13.6
5:	12,000	–	18,000	34.9	26.5
6:	18,000	–	24,000	46.1	39.3
7:	24,000	–	30,000	55.3	51.2
8:	30,000	–	36,000	64.7	62.3
9:	36,000	–	60,000	84.5	89.2
10:	60,000	–	90,000	93.7	97.6
11:	90,000	–	120,000	97.1	99.1
12:	120,000 +			100.0	100.0

Italy

Up to the income category 3 (3,600-6,000€) the household's income do not differ between ESS and ECHP8.

The cumulative responses up to category 6 (18,000 to 24,000€) differ about 2.5%. in ESS 63.9% of the households have an income up to 24,000€, in ECHP8 66.4% of the households are in the income categories 1 to 6.

In ESS, the median of the income measure is at class 6 and in ECHP8 the income median is the category 5 (12,000 to 18,000€).

In ECHP8 only 1.5% of the Italian households state a high income of 60,000€ and more, in ESS 4.5% of the respondents live in households telling this amount.

Taking the ECHP8 as a reference, interviewees of ESS with low or high household income overestimate the amount asked in the survey. Respondents in the middle categories of this monetary item underestimate the total net household income.

In general, we find small differences in the categorized measurement of household income between the two studies.

Luxembourg

In Luxembourg, lower categories of the income variable are not present in the wave 8 of ECHP. Only 0.2% of the households report an amount up to 6,000€ per year. The ESS tells us that 2.3% of the households are in the same income group.

In ESS, about 3% more households have income up to 24,000€; cumulative percent of all households from category 1 to 6 in ESS is 21% and in ECHP8 is 19%.

In the ESS of Luxembourg the median is at the income range of 30,000 to 36,000€, the median of ECHP8 is at the category 36,000 to 60,000€.

19.2% of the ESS respondents live in households with more than 60,000€. The same amount is given by 23.4% of the ECHP8 households.

Respondent with lower household income overestimate – and interviewees with high household income underestimate the amount of the total net household income during the ESS interview and compared to the ECHP8 outcomes.

In Luxembourg, the observed population with low income is rather small, whereas the upper end of the income distribution is common.

The upper half of the two cumulative frequencies shows remarkable differences in Luxembourg. In category 7 vary 9% points, in category 8 the difference is 12% points and in the ninth response category both surveys diverge with 4% points.

Table 4 Cumulative Frequencies of Total Net Household Income for Luxembourg

Income Categories	ESS	ECHP wave 8
1: up to 1,800	0.2	0.0
2: 1,800 – 3,600	1.3	0.1
3: 3,600 – 6,000	2.3	0.2
4: 6,000 – 12,000	3.5	1.3
5: 12,000 – 18,000	9.2	7.7
6: 18,000 – 24,000	21.2	19.4
7: 24,000 – 30,000	40.4	31.8
8: 30,000 – 36,000	54.6	42.6
9: 36,000 – 60,000	80.8	76.5
10: 60,000 – 90,000	94.1	93.9
11: 90,000 – 120,000	98.8	98.4
12: 120,000 +	100.0	100.0

The first comparative approach across the four countries shows a common pattern for the two data sets: The group of respondents with low household income (up to 6,000€ annually) and the respondent's group living in rich households (more than 60,000€) are rather small. In Germany, Italy, United Kingdom and Luxembourg the "poor" overestimate the total net household income. The "rich" also overreport the household's income in Germany, Italy, and United Kingdom, whereas in Luxembourg this group of people at the upper end of the income scale underestimates the amount. Taking the ECHP wave 8 as a reference, we see an underestimation of income in the middle categories of the income variable in every country. For Luxembourg the response rates in the central part of the income measurement are relatively close for the two studies used.

5 Characteristics Having an Impact on the Response Quality

From former research we assume that

1. the household size,
2. the selected respondent's knowledge about the financial situation of the other household members and the household as a total ,
3. the main source of incomes, and
4. the cognitive ability of the interviewee to remember the monetary amounts

will influence the response on total net household income.

5.1 The impact of household size

The ECHP8 is the 8th wave of a panel study, whereas the ESS is a strictly random sample with the known inclusion probability of each sampled unit into the survey.

In the lower income categories we find more households with one or two members. At the upper end of the income scale larger households are more frequent. This is true in all observed countries; and is much more noticeable in ECHP8 as in ESS. In greater households the probability increases to have more then one income earner. Having in mind, that an interviewed person does not like to answer in extreme responses, we assume that the respondent living in large households underreports the amount of the household income.

The ESS illustrates the following situation:

- In Germany one person households answer the question on household income with the lower income categories.
- In Italy and Luxembourg two- or three person households are positioned in the lower income groups.
- Large household with five or more members can also be found in the lower income classes of the ESS.

The ECHP8 points out the circumstances:

- In Italy, 3.9% of the low income households are units with five or more household members.
- In Germany and Luxembourg we do not identify large households in the lower part of the income distribution.

Both data show remarkable divergence of about 7 row % up to 14 row %. So far we conclude that the household income measurement of ESS is not reliable for research.

Table 5 Household Income Categories by Household Size in Germany, Italy and Luxembourg (row percent)

Income-Category	Germany				Italy				Luxembourg			
	Household size											
ESS	1	2	3,4	5+	1	2	3,4	5+	1	2	3,4	5+
1-3	60.9	24.1	8.0	7.0	23.2	37.5	30.4	8.9	27.3	22.7	36.4	13.6
4	55.7	26.4	15.1	2.8	17.9	32.5	36.6	13.0	66.7	16.7	8.3	8.3
5	39.8	36.1	21.2	2.9	9.8	27.6	53.7	8.9	36.4	16.4	32.7	14.6
6	13.0	61.9	31.1	3.9	9.5	24.8	60.0	5.7	35.0	29.9	28.2	6.9
7	8.6	37.6	43.2	10.6	5.7	19.5	64.3	10.3	18.2	24.6	44.9	12.3
8	6.9	36.1	51.6	5.4	6.7	15.6	51.1	26.6	13.0	28.3	46.4	12.3
9	7.2	38.6	46.4	7.8	1.4	10.1	71.0	17.4	8.3	18.5	59.4	13.7
10-12	7.8	35.8	43.0	13.4	6.9	3.4	69.0	20.7	2.1	21.9	55.1	20.9
ECHP8												
1-3	71.7	24.2	4.0	0.0	54.3	17.1	24.8	3.9	75.0	0.0	25.0	0.0
4	72.7	19.8	7.1	0.4	37.4	28.7	28.4	5.5	89.3	3.6	7.1	0.0
5	55.5	31.3	11.6	1.5	17.5	30.0	44.1	8.3	70.3	20.0	9.0	0.6
6	22.9	48.0	25.1	4.0	5.2	29.2	56.1	9.4	53.7	30.7	13.7	1.8
7	8.8	39.1	44.0	8.0	2.4	17.0	69.1	11.5	35.9	36.9	23.6	3.7
8	4.4	36.4	51.5	7.8	1.6	16.9	68.3	13.2	25.4	37.5	29.5	7.6
9	2.7	31.8	54.1	11.4	2.0	9.9	65.7	22.3	9.8	34.3	45.4	10.5
10-12	6.3	24.4	52.0	17.2	6.2	17.3	63.0	13.6	3.9	27.0	53.0	16.1

5.2 The impact of the respondent's family relation to the main income earner

The ESS sample design selects randomly a household member as interview partner. A responding person can have a close family relationship to the main income earner. These are the partner of the main bread winner and him or her self. The other cases like the children and/or the parents and/or other relatives we interpret as interviewees, having a distant relation. During the interview we expect that answers form a close respondent are more reliable then from a far-away person. By increasing distance to the main income earner, the answer underestimate the "real" total net household income, because the state of information about the financial situation of the household decreases.

Table 6 Household Income Categories by Respondent's Relation to the Main Income Earner in Germany, United Kingdom, Italy and Luxembourg (cumulative percent) in the ESS

Income Category	Germany		United Kingdom		Italy		Luxembourg	
	Relation to Main Income Earner							
	close	distant	close	distant	close	distant	close	distant
1	0.2	1.3	0.5	1.4	0.9	0.5	0.3	0.0
2	0.9	3.3	1.1	4.9	2.9	2.6	0.9	2.2
3	1.6	8.8	3.2	10.4	7.4	12.0	1.7	3.4
4	6.3	28.0	13.5	36.1	24.9	35.4	2.2	6.2
5	19.0	53.2	24.9	50.6	44.3	54.7	5.7	16.1
6	39.8	65.8	36.9	60.7	62.7	66.7	14.8	34.2
7	59.4	76.6	47.1	68.2	77.8	77.1	34.2	53.1
8	73.2	83.9	57.1	76.6	84.7	84.4	48.6	66.8
9	91.2	95.1	81.0	89.9	95.5	95.3	77.4	87.6
10	97.4	98.7	92.5	95.7	99.1	97.9	92.3	97.8
11	99.1	99.1	96.7	97.7	99.6	99.0	98.5	99.4
12	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
valid n	1,640	696	1,092	692	445	192	650	322

The table shows that far-away replaying persons answer the income questions by ticking one or two income categories lower than the main income earner or his/her partner.

5.3 The impact of the main source of income

Other sources of inaccuracies in the measurement of income are the main sources of income. A respondent living in a household with the income mainly from work can remember the periodical and regular amount of the wage and salary. The same is true for pensions as a main source. Unemployment benefits, social benefits or grants, income from investments, savings or property and income from other sources are additional elements, which the respondent has to add-up to the total net household income. An increasing number of income sources will increase the complexity of remembering and summing up the household income. Particular difficulties to answer the income question we expect from respondents living in households with self employment income as the main source.

Table 7 Main Source of Household Income (column percent)

Main Source	Germany	United Kingdom	Italy	Luxem-bourg
ESS				
Wages and Salaries	58.1	57.5	57.2	63.7
Income from self-employment or farming	6.6	4.3	16.8	6.8
Pensions	26.4	26.3	23.5	26.0
Unemployment, redundancy benefit	4.5	1.7	0.9	0.9
Any other social				
Any other social benefits or grants	2.0	8.1	0.6	1.3
Income from investments, savings, etc.	0.6	1.0	0.2	0.1
Income from other sources	1.8	1.1	0.8	1.1
valid n	2,893	2,029	1,123	1,510
ECHP8				
Wages and Salaries	61.6	58.6	49.5	65.0
Income from self-employment or farming	5.4	5.7	15.2	3.0
Pensions	23.9	23.2	30.2	24.8
Unemployment, redundancy benefit	3.0	0.3	1.0	0.2
Any other social benefits or grants	4.2	9.8	2.0	5.9
Private income	1.9	2.4	2.0	1.2
valid n	5,559	4,779	5,525	2,428

Both surveys give the same impression on the main income sources of the households.

In Germany, United Kingdom and Luxembourg the most frequent monetary resource is income from dependent work, followed by pensions and retirement benefits. Both categories cover 80% to 90% of all income sources.

In Italy the ESS reports a lower rate of pensioners as the ECHP8, the ESS surveyed 23.5% households with old age pensions and the ECHP8 reports that 30.2% of the Italian households have pensions as the main income source.

In Germany we also see a remarkable proportion of household living from unemployment benefits.

In United Kingdom social transfers are often given as main income source (9.6% of the ECHP8 households and 8.1% of the ESS households).

In Luxembourg, the respondent from a household with self-employment income as main source underreports the amounts in ESS compared to the ECHP wave 8.

In Germany and Italy, the highest income category of self-employed is overestimated during the interviews of ESS.

Wages and salaries and pensions show in both surveys the similar answering behavior of the interviewees.

Comparing social transfers in ESS and ECHP8, only very few cases are observed in Luxembourg and Italy who answered the income measurement and social benefits as the main source. In ESS, about 2/3 of the respondents with social transfers ticked the lowest income categories; in ECHP8 only 28% of the households have less then 12,000€ annual total net income.

5.4 The impact of income composition

The ECHP interviews ask for 21 possible income sources. Every member of a household aged 15 and older is requested to remember these monetary items and give the amount received.

Table 8 Number of Income Sources by Proportion of Individuals in ECHP Wave 8

	Germany	United Kingdom	Italy	Luxembourg
no income source	6.6	1.4	24.6	17.3
one income source	0.0	0.0	0.0	0.0
2	0.8	0.6	1.1	0.0
3	5.5	5.3	11.5	7.0
4	7.6	5.9	17.7	10.5
5	5.3	8.8	6.3	26.1
6	19.8	25.6	27.1	8.6
7	18.2	12.4	3.3	19.7
8	9.5	18.2	6.1	4.1
9	9.4	11.1	1.7	4.6
10	7.0	5.8	0.4	1.9
11	8.6	3.6	0.2	0.3
12	1.5	1.0	0.0	0.0
13 and more income sources	0.2	0.2	0.0	0.0
valid n	10,624	8,521	13,392	4,916

Most of the persons have to give an account for five or six different incomes.

In Italy 24% of the ECHP individuals have no income from any source. The highest proportion of people having income receive the money from six various sources. 63% of the Italians have three to six different incomes.

In Germany between six and eleven income sources are answered. More then 72% of the individual have to report on such complex income composition.

In the United Kingdom most of the interviewees have to remember five to nine sources of revenues. 9% of the ECHP individuals have more then nine income sources.

In Luxembourg, most people have to sum up five different income components, and 17% have no income sources to mention. Only 11% of the Luxembourg ECHP individuals have more then seven different incomes.

5.5 The impact of remembering income

The detailed fieldwork instrument of ECHP shows the complexity of to measure total net household income. At least four different sources, in average six and sometimes 13 and more income components are reality for the respondent.

The straightforward questions of ESS recall only the main income source of the respondent's household. These are income from work, a periodical source and a constant amount of money, the interviewed person can answer the ESS query. The same is true for payments replacing the income from work, like pensions, unemployment benefits and alimonies; these are easily remembered by the interviewees.

For all other types of income the questionnaire has to ask separate questions to remind the interview partner about this monetary item.

At the same time, the household member selected for the interview must have the knowledge about the variety of the household income components. The ESS surveyed a randomly selected member of the household as a reference person. This can be the main income earner or his/her partner with a good knowledge on the income or other household members having weak information about monetary items.

The following graphs illustrate the proportion of well informed respondents minus the proportion of less informed interviewees by income category. A negative bar shows that more less informed interview partners then well informed have chosen that income brackets.

The less informed reference persons dominate in the lower income categories. In Germany, the impact on the fourth and fifth income group is observable. In United Kingdom, the less informed persons of contact have an influence only on category 5; up to the income group 8, there is a balance between good informed answers and reference persons with a weak knowledge on the total net household income.

For Italy, we assume that in category 3 and 4 the less informed people underestimate the amount of the household income, and there is a slight effect on the top two income groups.

In Luxembourg, the influence of respondents with less knowledge on the total household income is visible in the lower part of the income distribution.

For the other states of ESS, we observe that up to the income category 8 in countries with an higher average of total net household income the proportion of less informed respondents are larger then the proportion of well informed; and we again assume that the sum of the total net household income is underestimated (e.g. Switzerland, Sweden and Finland). The impact of less informed reference persons in countries with a lower average of income is seen in the categories 1 and 2; in Portugal, Hungary and Poland these income ranges are dominated by the less informed answering person.

A particular situation is empirical visible in Poland. From category 5 to category 11 we have as much informed as not informed responses and the twelfth group is mainly built by respondents with less knowledge of the income.

Figure 4 Well Informed vs. Less Informed Interviewees in Germany

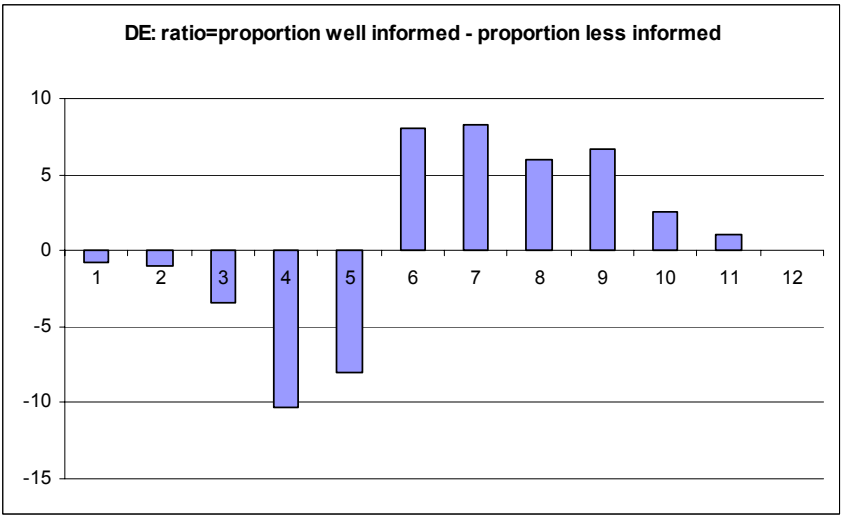
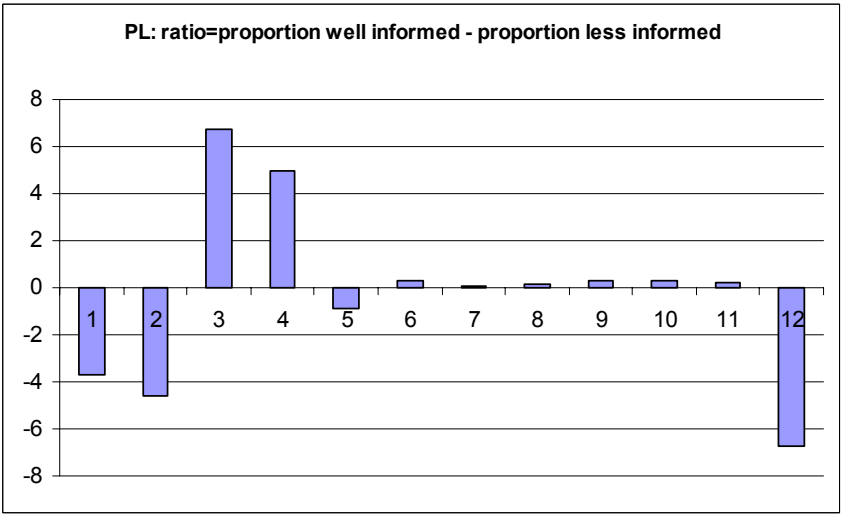


Figure 5 Well Informed vs. Less Informed Interviewees in Poland



6 Categorizing Income for Comparative Social Research

We are looking for "optimal" answer categories for the interviews asking the income question in various national contexts.

By cutting the income variable of ECHP8 into 5% groups of the population and sorting the ESS categories into the ECHP8 distribution, we illustrate the need to adjust the income brackets to national financial circumstances and the national income distributions.

The ESS category 36,000€ to 60,000€ covers the 9th to the 15th 5% percentiles of the income distribution in Luxembourg. In Germany, the same income group covers the 15th to 19th 5% percentiles. In Portugal, the richest 5% of the population have a total net household income of 36,000€ to 60,000€. Also, the poorest 5% of the Luxembourg people have a higher household income than 55% of the Portuguese population and 50% of the Italians.

Respondents from all countries need about six ESS categories to answer the income question.

We propose for Germany, United Kingdom and Finland a system of income categories starting with an annual total net household income up to 5,000€. The scale continues in 5,000€ steps to the amount of 60,000€. The top category is 70,000€ and more (see Tables 10 and 11).

Table 9 5% Percentiles of the Total Household Net Income in ECHP8 for Selected Countries

Income Percentiles No	%	Germany	United Kingdom	Italy	Luxem-bourg	Portugal	Finland
1	5%	8,658	7,781	5,163	16,039	2,394	6,203
2	10%	11,327	10,632	7,218	19,503	3,328	8,309
3	15%	13,752	12,535	8,728	22,310	4,141	10,258
4	20%	15,769	14,961	10,071	24,374	4,920	12,504
5	25%	17,507	17,271	11,310	27,088	5,658	14,504
6	30%	19,537	19,612	12,395	29,509	6,453	16,176
7	35%	21,249	21,829	13,634	32,308	7,388	17,844
8	40%	23,129	24,316	14,901	34,620	8,394	19,654
9	45%	24,745	26,774	16,205	37,067	9,389	21,432
10	50%	26,541	29,400	17,849	39,530	10,385	23,572
11	55%	28,032	31,865	19,419	42,142	11,333	25,765
12	60%	29,780	34,816	21,156	45,378	12,381	28,056
13	65%	31,767	37,552	22,987	49,571	13,553	30,226
14	70%	33,816	40,861	25,100	53,859	14,816	32,438
15	75%	36,108	44,335	27,165	59,059	16,398	34,883
16	80%	39,097	48,239	29,541	63,653	18,516	37,697
17	85%	42,763	53,432	32,592	70,746	20,950	40,990
18	90%	47,796	61,142	37,092	79,787	24,744	46,582
19	95%	56,613	72,806	45,489	95,240	32,166	56,414
Valid N		5,559	4,779	5,525	2,428	4,588	3,106

In Luxembourg the income responses begin with the income up to 10,000€. At the top of the income scale Luxembourg needs 10,000€ brackets until 110,000€ is reached.

Italy and Portugal need an extension at the bottom part of the income distribution. The first group is the annual household income up to 2,500€, continued in 2,500€ classes until 15,000€ is reached. From here, 5,000€ groups up to the top of 60,000€ completes the income response categories.

The proposed income categories take into account the differences in the national income distributions (see Table 11). These diversities are observed and measured by income brackets of 5000€. For countries with a larger population at the bottom end of the income curve, the income classes are in 2500€. At the top end of the income inequality, our proposed income scales take into account the population size with high incomes. In a wealthy country, the scale continues in 10000€ brackets. Comparing Luxembourg and Portugal illustrates the advantages. 0.5% of the population in Portugal has a total net household income of 60,000€ and more; but every fourth respondent in Luxembourg lives household with 60,000€ and more.

Table 10 The Distribution of the Nineteen 5% Percentiles from ECHP8 by the 12 Income Categories of ESS in Selected Countries

ESS categories	Germany	United Kingdom	Italy	Luxem bourg	Portugal	Finland
	No. of the ECHP8 5% percentile					
up to 1,800	---	---	---	---	---	---
1,800 – 3,600	---	---	---	---	1-2	---
3,600 – 6,000	---	---	1	---	3-5	---
6,000 – 12,000	1-2	1-2	2-5	---	6-11	1-3
12,000 – 18,000	3-5	3-5	6-10	1	12-15	4-7
18,000 – 24,000	6-8	6-7	11-13	2-3	16-17	8-10
24,000 – 30,000	9-12	8-10	14-16	4-6	18	11-12
30,000 – 36,000	13-14	11-12	17	7-8	19	13-15
36,000 – 60,000	15-19	13-17	18-19	9-15	---	16-19
60,000 – 90,000	---	18-19	---	16-18	---	---
90,000 –120,000	---	---	---	19	---	---
120,000 and more	---	---	---	---	---	---

Table 11 Proposed System of Income Categories for European Countries

Income Categories No	in EURO	Germany	United Kingdom	Italy	Luxem-bourg	Portugal	Finland
1.0	up to 2,500			1.5		5.8	
2.1	2,500 – 5,000			3.8		16.3	
2.2	up to 5,000	0.9	1.9				3.5
3.0	5,000 – 10,000	6.2	7.8				10.9
3.1	5,000 – 7,500			7.3		15.7	
3.2	7,500 – 10,000			9.3		13.1	
3.3	up to 10,000				0.6		
4.0	10,000 – 15,000	11.1	12.3		3.2		12.3
4.1	10,000 – 12,500			11.8		12.0	
4.2	12,500 – 15,000			10.3		10.1	
5	15,000 – 20,000	13.6	11.7	16.6	7.1	11.6	15.0
6	20,000 – 25,000	15.1	10.3	12.9	9.9	7.0	11.9
7	25,000 – 30,000	15.1	10.1	10.4	10.9	3.4	11.4
8	30,000 – 35,000	12.2	8.7	6.5	8.9	1.9	10.6
9	35,000 – 40,000	8.6	8.3	3.9	10.0	1.0	8.5
10	40,000 – 45,000	5.6	6.3	1.7	8.9	0.7	4.8
11	45,000 – 50,000	4.2	6.0	1.6	6.3	0.3	3.5
12	50,000 – 55,000	2.5	4.1	0.8	5.5	0.3	2.2
13	55,000 – 60,000	1.3	3.0	0.7	5.2	0.3	1.4
14	60,000 and more			0.9		0.5	
15	60,000 – 70,000	1.7	4.3		8.1		1.8
16	70,000 and more	2.0	5.3				2.2
17	70,000 – 80,000				5.5		
18	80,000 – 90,000				3.9		
19	90,000 – 100,000				2.5		
20	100,000 – 110,000				1.2		
21	110,000 and more				2.3		

References

European Commission. 1996. *European Community Household Panel (ECHP): Volume 1. Survey Methodology and Implementation.* Luxembourg: Survey Questionnaires.

European Commission/Eurostat. 2003. *ECHP UDB Description of Variables. Data Dictionary, Codebook and Differences between Countries and Waves.* (DOC.PAN 166/2003-12). Luxembourg.

European Commission/Eurostat. 2003. *ECHP UDB Construction of variables. From ECHP questions to UDB variables.* (DOC.PAN 167 / 2003-12). Luxembourg.

European Social Survey 2002. *The European Social Survey.* Source Questionnaire (Round 1, 2002). ESS Document Date. 01-08-02.

European Social Survey. 2002. *The European Social Survey.* Source Showcards.

European Social Survey. 2002. *The European Social Survey. Project Instructions (PAPI).* ESS Document Date. 15-07-02.

Hoffmeyer-Zlotnik, J. H.P., and U. Warner. 1998. „Die Messung von Einkommen im nationalen und internationalen Vergleich." *ZUMA-Nachrichten* 42:30-68.

Hoffmeyer-Zlotnik, J. H.P., and C. Wolf (Eds.). 2003. *Advances in Cross-National Comparison. A European Working-Book for Demographic and Socio-Economic Variables.* New York: Kluwer Academic/Plenum Publisher.

Lyberg, L., P. Biemer, M. Collins, et al. 1997. *Survey Measurement and Process Quality.* New York: Willey & Sons.

Warner, U., and J. H.P. Hoffmeyer-Zlotnik. 2003. "How to Measure Income." Pp. 307-323 in *Advances in Cross-National Comparison. A European Working-Book for Demographic and Socio-Economic Variables,* edited by Hoffmeyer-Zlotnik, J. H.P., and C. Wolf. New York: Kluwer Academic/Plenum Publisher.

HOW TO MEASURE EDUCATION IN CROSS-NATIONAL COMPARISON: HOFFMEYER-ZLOTNIK / WARNER-MATRIX OF EDUCATION AS A NEW INSTRUMENT

JÜRGEN H.P. HOFFMEYER-ZLOTNIK & UWE WARNER

The comparative measurement of education is a complex task. The national systems of education and schooling are differently organized across national states and nations. In this paper we will sort the certificates from general and professional schools into one matrix that allows us to compare the "highest level of education obtained" across countries.

1 The Problems

Because of their historical development and their political tradition, national education systems are particular for each nation. In general, each school system incorporates in general education the pre school and the basic school education with a various number of degrees to obtain; in the professional education with the whole range between school based and vocational, enterprise based training and all the possible mixtures; and finally in high school education with its entire spectrum of diplomas. Common to all are four sections:

- The primary section, including the pre school and basic education for 4 or 6 years of schooling;
- the lower secondary programs cover in most European countries the general education until the end of basic education with a first school certificate after 8 to 11 years of schooling;
- the upper secondary segment includes the school institutions until the entry to high school, and the professional training until the first vocational certificate that allows to execute the learned profession, but lower then high school degrees;
- the tertiary section contains all the different types of high schools, the applied universities and the universities with the academic education until research qualifications are obtained.

So far, three common anchor points can be identified: the basic certificate which differs across countries by duration of schooling and the pupil's age, the highest possible degree of general education as the entry point to university and in general obtained after 12 or 13 years of schooling, and finally the end of university education with the PhD thesis.

The differences across the national education systems are based on various objectives about the optimal function and the aims of education. The institutionalization of schooling is driven by national ideologies and traditional developments, and education is finally codified in national law.

The definition of "basic education" varies across the countries. And the meaning of "basics" has an impact on the duration of schooling for a basic degree, the description of compulsory full-time school, the differentiation into parallel types of school and the split off point before (e.g. in Germany) or after (e.g. in Denmark) the basic degrees, and with the impact on the parent's decision for further education for their children. Characteristics influenced by the definition of basic education are the national structures of the school institutions: Are there degrees depending on each other, degrees increasing in their valid and built on upon as sequences of educational careers? Beside this vertical structure are there horizontal differentiations of educational institutions, and is it and how difficult is it to switch from one track to a parallel path upwards? How permeable are the national types of school?

The national education systems are also structured by

- the (legal) rules on entry and leaves to dedicated school types and levels,
- the duration of minimum and maximum schooling periods,
- the possibilities to repeat classes, and
- the maximum number of allowed repetitions.

An important factor is the degree of side by side existence of private and public schooling in the general and professional training sectors. Of course, the transition from general to vocational sectors is characterizing the national school system. The differentiation of professional education certificates and their following up rules are of importance. An important question concerns the political and social acceptance of schools and their diploma as well as the legal and political control of the state.

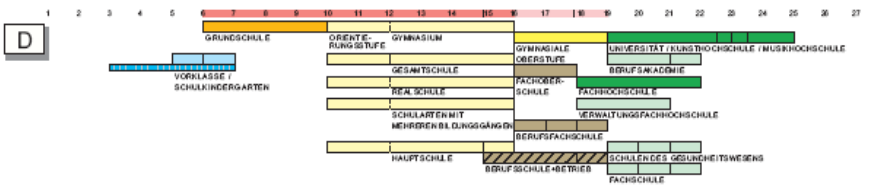
2 The National Education Systems

This chapter describes the school systems of three European countries: Germany, Denmark and Luxembourg. In main and fundamental points they differ. The following chapter introduces the usual categories of cross country comparison and in the next part we illustrate the Hoffmeyer-Zlotnik/Warner matrix that compares the national education systems.

2.1 Education in Germany

In Germany, compulsory education lasts for 9 school years. From their 6 year of age onwards children attend "Grundschule" for 4 classes. After this primary part they can choose at least between three types of secondary schools: "Hauptschule" for the next 5 school years, the 6 school years long "Realschule", or "Gymnasium" for the next 9 classes.

After leaving "Hauptschule", having finished the lower secondary education, vocational training in the dual system or in vocational school is possible and became the normal school career. After finishing "Realschule" it is possible to continue with "Fachoberschule". "Abitur", the degree obtained in "Gymnasium", is the standard entrance diploma to university and finishes upper secondary education.



Source: European Commission, 2002

Because of the paralleled general and vocational education, in German research we have to ask two interview questions about education during social surveys.

1. The *general education* with three or five types of lower secondary school (depending on method of counting) and two degrees of upper secondary school certificates.
2. The *vocational education* with answer categories for the dual system and for professional schools, for different types of schools and answer possibilities for vocational or technical full time school degrees and vocational colleges, and categories for university diploma.

Therefore, German social survey research needs a two dimensional matrix for the construction of a rank order concerning educational attainment or a hierarchical social order of educational levels.

Table 1 General Education by Vocational Education, *Germany*, ESS 1st Round

general education by degree	vocational education by degree							total
	non	dual system	vocational-school	vocational college	univ.of applied sciences	university	others	col %
non	14,3	1,4	,7	,0	,0	,3	1,4	2,2
8th/9th class	64,8	49,2	31,6	27,0	3,8	1,4	28,2	37,4
10th class	11,5	42,2	46,3	49,2	24,6	2,4	52,1	34,9
restricted Abitur	,8	2,4	8,8	11,8	27,7	7,4	5,6	6,2
Abitur*)	7,0	4,7	11,8	11,5	41,5	86,1	9,9	18,5
others	1,6	,2	,7	,5	2,3	2,4	2,8	,9
row %	10,1	48,0	5,6	15,8	5,4	12,2	2,9	100,0
total	244	1161	136	382	130	296	71	2420

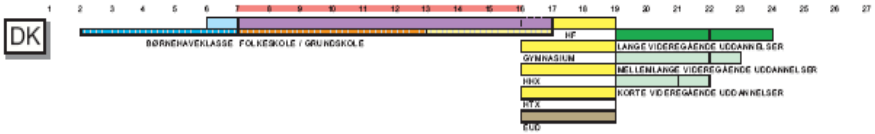
*) University-entrance diploma
Source: ESS round 1, computation by the authors

In Germany, education is a central policy issue of the 16 “Bundesländer” (federal states). Each state manages his own educational system and the “Bundesländer” agreed on the transfer between school types, the recognition of qualifications and the entrance possibilities to the further schools. Table 1 already summarizes the systems of the16 “Bundesländer” and presents their common structure.

2.2 Education in Denmark

In Denmark, compulsory education starts at the age of 6 at "Folkeskole" and lasts for all pupils for 9 years (as comprehensive school). A voluntary 10th year, or the Gymnasium (for 3 years), or vocational education follows.

The general upper secondary education is much diversified as in Germany, whereas the primary and lower secondary sectors are unified into one track of schooling and the tertiary sector offers three types of high schools.



Source: European Commission, 2002

Table 2 Highest Level of Education, *Denmark*, ESS 1st Round

Categories	total	Valid Percent
0 No school education, no vocational education	2	,1
1 1.-6. class in school, no vocational education	18	1,2
2 7.-10. class in school, no vocational education	351	23,5
3 Upper secondary school, no vocational education	103	6,9
4 Vocational education and training, apprenticeship training	594	39,8
5 Work leader education for vocational educated	32	2,1
6 Further education of 2-3 years after upper secondary school	137	9,2
7 Further education of around 4 years after upper secondary sector	149	10,0
8 Bachelors or masters degree from university	98	6,6
9 Further university education i.e. Ph.D	10	,7
Total	1494	100,0

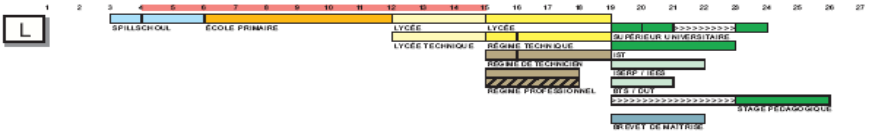
Source: ESS round 1, computation by the authors

In Denmark, the ESS surveys highest level of education by a 10 category answer scheme of school leaving qualifications. They look already being created in advance for the recodes into the International Standard Classification of Education (ISCED 1997) demanded by the coordinators of ESS.

2.3 Education in Luxembourg

In Luxembourg, the primary school starts at the age of 6 and ends at the age of 12. The secondary sector is divided into complementary, technical and general schools. The duration of "lycee" varies between 3 and 7 classes.

Several vocational schools and a university of applied sciences do also exist. The upper secondary education is very diverse and the third sector contains several professional educational institutions.



Source: European Commission, 2002

Table 3 Highest Level of Education, Luxembourg, ESS 1st Round

Categories		total	Valid Percent
0	Pas de diplôme/qualifications	20	1.3
1	Ecole primaire	254	16.7
2	Primaire supérieur	120	7.9
3	Enseignement complémentaire	98	6.4
4	Certificat d'enseignement secondaire technique inférieur	52	3.4
5	Certificat d'apprentissage	22	1.4
6	**C**ertificat de **C**apacité **M**anuelle	22	1.4
7	**C**ertificat d'**I**nitiation **T**echnique et **P**rofessionnelle :	36	2.4
8	**C**ertificat d'**A**ptitude **T**echnique et **P**rofessionnelle :	237	15.6
9	Diplôme de technicien (jusque 13e dans le régime tech.)	36	2.4
10	Bac technique (jusque 13e ou 14e du régime technique)	50	3.3
11	Enseignement secondaire général inférieur	115	7.6
12	Diplôme de fin d'études secondaires	139	9.1
13	Brevet de maîtrise artisanale	32	2.1
14	Enseignement supérieur - BAC +2	53	3.5
15	Enseignement supérieur - BAC +3	57	3.7
16	Enseignement supérieur - BAC +4	57	3.7
17	Enseignement supérieur - BAC +5 ou plus	57	3.7
18	Enseignement supérieur - Doctorat	11	.7
19	Autre: Précisez	43	2.8
Total		1523	100.0

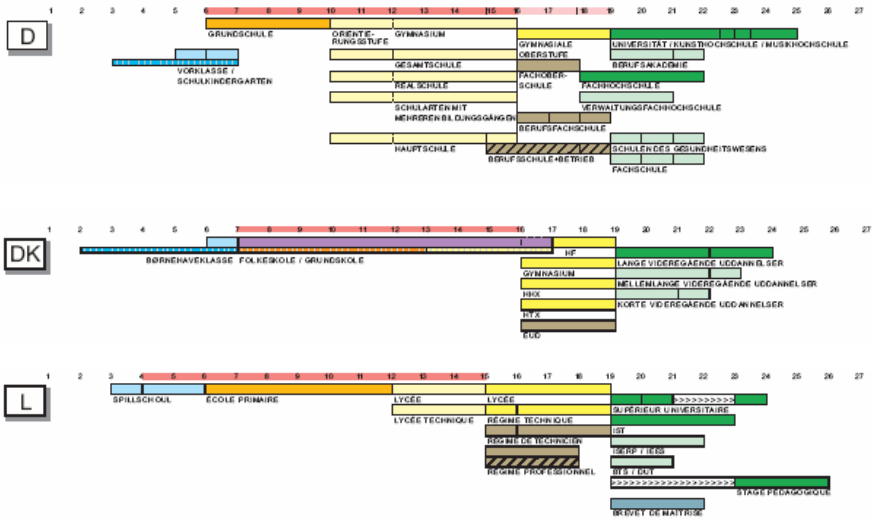
Source: ESS round 1, computation by the authors

In Luxembourg, the fieldwork for ESS used 19 different answer categories to obtain the information about the highest level of education. On a first glance, the proposed certificates are much more detailed than in Germany and Denmark and they do not summarize the national education system. The labor market in Luxembourg is characterized by a very high proportion of non-Luxembourgish employees and workers who are not educated and trained in the national education system. Therefore the response categories of the survey question on highest level of education must also cover qualifications obtained in the neighboring countries of Luxembourg.

2.4 Comparison of the educational institutions

While comparing the educational systems across the three countries we can identify

- that in Germany, the differentiation into the educational tracks starts after 4 classes in primary school. In Denmark, this separation takes place after 10 school years; in Luxembourg after 6 years of primary classes;
- that in Germany and Luxembourg the tertiary education is more differentiated than in Denmark; and
- that among the studied countries, only Germany includes the dual system combining school education and training organized at the workplace inside the enterprises. In the other two countries the vocational education takes place in (particular) professional schools.



Source: European Commission, 2002

3 Measurement Instruments for Cross-National Comparison

Today, in comparative research four instruments measuring and comparing highest level of education can be identified (Braun & Müller, 1997; Hoffmeyer-Zlotnik & Wolf, 2003):

- years of schooling;
- the "International Standard Classification of Education" (ISCED 1997)
- the "CASMIN Educational Classification"; and
- the "Hoffmeyer-Zlotnik Educational Classification".

3.1 Years of schooling

In surveys for cross-country comparison the instrument "years of schooling" is the most used one for the measurement. But various surveys use different questions and wordings and focus on the information in slightly different manners:

- the *European Social Survey (ESS)*, round 1, question F7 asks: "How many years of full-time education have you completed?";
- the *International Social Survey Programme (ISSP)* is asking about "years (of full time) schooling including university but not vocational training";
- the *General Social Survey (GSS)* of the U.S. asks about "grades" and "years of college" (NORC and Roper, 1996: 49); and
- the German *Sozialwissenschaften-Bus* 1996 (social science bus survey) question wording is: "In which age you left general school?" (GFM-GETAS/WBA, 1996: 2).

All four questions generate different answers. ESS and ISSP obtain the number of years spent in educational institutions, and the ISSP does not include years spent in vocational education. The question about years only makes sense in cases where the repetition of classes is not foreseen and allowed. In this case a question about grades like in the American GSS produces the informative measure. The German social science bus survey asks about the age when the respondent left school; but leaving school at an older age does not necessarily lead to a higher degree of education.

3.2 International Standard Classification of Education – ISCED 1997

The "International Standard Classification of Education – ISCED", (UNESCO, 1997, 2003) was developed in the seventies by UNESCO. The major aim was to unify international statistics on educational levels of the population. The actual version of this classification was revised in 1997 and offers a common set of concepts, definitions and classifications establishing a frame for collecting data and presenting indicators on outcomes of the

school systems. It covers all teaching activities organized in educational institutions for pupils and adults from pre school education to continued schooling and training as well as general and vocational education. Seven categories are offered by this classification.

Table 4 International Standard Classification of Education – ISCED 1997

Name of the Level	Code	Complementary Dimensions
Pre-primary education	0	None
Primary education, First stage of basic education	1	None
Lower secondary education, Second stage of basic education	2	Type of subsequent education or destination, Programme orientation
(Upper) secondary education	3	Type of subsequent education/destination, Programme orientation, Cumulative duration since beginning of ISCED level 3
Post-secondary non tertiary education	4	Type of subsequent education/destination, Cumulative duration since beginning of ISCED level 3, Programme orientation
First stage of tertiary education (not leading directly to an advanced research qualification)	5	Type of programmes, Cumulative theoretical duration at tertiary, National degree and qualification structure
Second stage of tertiary education (leading to an advanced research qualification)	6	None

see: UNESCO, 2003: 203

3.3 The CASMIN Educational Classification

The CASMIN Educational Classification "distinguishes educational levels according to their selectivity effects. In this respect, the schema claims *functional equivalence* of its educational categories across countries. The criterion of selectivity combines two perspectives: demarcation of typical class-barriers in the educational system on the one hand, and identification of decisive signals for utilisation on the labour market on the other. Following these considerations, the CASMIN schema is constructed as a *certificate-oriented* classification" (Brauns, Scherer & Steinmann, 2003: 222).

The CASMIN Educational Classification is a hierarchically structured measurement of certificates and is two dimensionally separated into general and vocational qualifications. This classification is also based on the institutional structure of educational sectors and divides the secondary part into three hierarchical steps and the tertiary sector into two sub categories of professional orientation and academic degrees.

Table 5 The CASMIN Educational Classification

Level		CASMIN	Description	
Tertiary	High	3b	Higher tertiary education: The completion of a traditional, academically-oriented university education	
	Low	3a	Lower tertiary education: Lower-level tertiary degrees, generally of shorter duration and with a vocational orientation	
Secondary	High	2c_voc	Vocational maturity: Full maturity certificates including vocationally-specific schooling or training	
		2c_gen	General maturity: Full maturity certificates (e.g. the Abitur, A-levels)	
	Mediate	2a voc	Intermediate vocational qualification, or secondary programmes in which general intermediate schooling is combined by vocational training	
		2b gen	Intermediate general education Academic or general tracks at the secondary intermediate level	
	Low	1c voc	Basic vocational training above and beyond compulsory schooling	
		1b gen	General elementary education	Social minimum of education. It generally corresponds to the level of compulsory education
Primary		1a	Inadequately completed general education	

voc=vocational education, gen=general education

Brauns, Scherer & Steinmann, 2003: 223

3.4 Hoffmeyer-Zlotnik Educational Classification

The classification proposed by Hoffmeyer-Zlotnik Educational Classification (Hoffmeyer-Zlotnik, 2003) is also based on school leaving qualifications. He combines certificates from general and professional education. Having in mind which average occupational prestige a respondent can obtain on the labor market by this combination of certificates, he rank orders the categories. He uses the Standard International Occupational Prestige Scale (SIOPS) developed by Treiman (1977; Ganzeboom & Treiman, 2003). SIOPS derives from the International Standard Classification of Occupations (ISCO 88) and measures the professional activity of an observed respondent. Hoffmeyer-Zlotnik's main argumentation is that for executing a profession a social recognized qualification is necessary and of central importance. This obtained qualification leads to a corresponding amount of social reputation as long as the educational institutions are controlled by the state and the

achievement of a diploma is required for exercising that profession. Combining educational outcomes and the occupational activity is (at least for modern societies) important, because the accreditation of occupational carriers depends on the finished educational background.

This classification does not distinguish between sectors of education but does cover the various combinations of general and vocational degrees. It allows an overview on the entity of certificates in a studied country. Table 6 illustrates the relation between general and vocational education and the average prestige scores of German respondents.

Table 6 Hoffmeyer-Zlotnik Educational Classification Demonstrated at the Case of *Germany*

Code	General Education	Vocational Education	average occupational Treiman prestige
1	no basic degree	none	14-20
2	basic degree	none/unfinished	15-20
3	no basic degree	operational	20-30
4	basic degree	operational	20-35
5	basic degree	vocational school	20-35
6	middle degree	none/unfinished	20-35
7	middle degree	operational	25-35
8	middle degree	vocational school	25-45
9	higher degree	operational	30-40
10	higher degree	vocational school	40-55
11	middle degree	vocational college	50-65
12	higher degree	technical college	50-70
13	higher degree	university, 1st degree, BA	65-75
14	higher degree	university, 2nd degree, MA	70-78
15	higher degree	university, doctorate, Dr./Ph D	70-78

Hoffmeyer-Zlotnik, 2003: 254

4 Problems of Misclassification

The first problems in comparing "highest level of education" are demonstrated in the chapter on "years of schooling" as comparative measurement. Different question wording in the surveys creates different information substituted under similar variable labels. The evaluation of this item across the surveys becomes impossible, because different fieldwork instruments produce diverse measures. Even if the questionnaire offers a number of answer categories like ISCED 1997, the researcher can face unsolvable problems in comparing countries or educational systems. The common standards differ from national accepted customs and habits. Sorting the national degrees and certificates, the reclassifica-

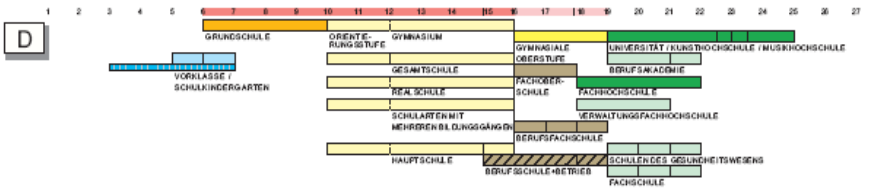
tion of the country's educational systems bears a serious degree of freedom for the researcher. Table 7 confronts how the German ESS national coordination team classified the survey outcomes into the common standards of the ESS coding and the author's exercise based on the ISCED 1997 instructions and manual. The main variation is detected for the category "primary or first stage of basic". ESS reports 1.7% of the respondents having obtained this degree. But the German educational system does not allow leaving the school institutions at that grade. Therefore only 0.1% can be found in our regrouping. The class "post secondary, non tertiary" also varies. The differences are due to the degree master craftsmen (Handwerksmeister) and due to respondents having done obligatory practical courses and internships after having reached the university entrance diploma and the start of university education.

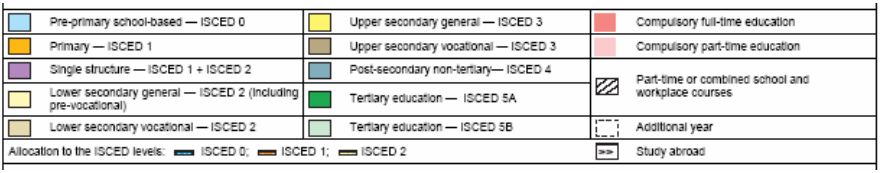
Table 7 Misclassification – The case of ISCED 1997 categories for *Germany*

	ISCED Categories for Germany	ESS classification	our reclassification of ESS
1	Primary or first stage of basic	1,7	0,1
2	Lower secondary or 2nd stage of basic	13,8	13,4
3	Upper secondary	58,2	54,4
4	Post secondary, non-tertiary	4,8	7,1
5	First stage of tertiary	20,5	23,5
6	Second stage of tertiary	1,1	1,1
N		2916	2906

Source: ESS round 1, computation by the authors

Particularly vocational degrees in the German educational system are not easy to classify by ISCED 1997. CEDEFOP, the European Center for the Development of Vocational Training categorized the German educational system by ISCED 1997 without using category 4: "post secondary, non tertiary". But for the degree master craftsmen (Handwerksmeister) there is no other possibility as to categorize these degrees in ISCED category 4.





Source: European Commission, 2002

5 A Proposal for Level of Highest Education Based on a Matrix with 10 Categories

5.1 Building the Hoffmeyer-Zlotnik/Warner Matrix of Education

The Hoffmeyer-Zlotnik/Warner Matrix of Education has the advantage to minimize the errors of misclassifications described above.

The matrix is built on the answers to the interview question on the highest general educational level obtained and the vocational education degree. One dimension presents the general education and the other axis the professional education including high school and university diploma. All national possible degrees relevant in the national education system are rank ordered from not applicable, lowest level to highest certificate.

The second step for creating the matrix is to bring the combination from general and vocational degree together with the social prestige that a person can gain on the labor market. The prestige scores are also ranked from low to high. Grouping together combinations of degrees with the similar prestige we come up with 10 valid categories and the 0 represents combinations not possible in the national system of education.

Table 8 Hoffmeyer-Zlotnik/Warner Matrix of Education – for *Germany*

vocational education	general education				
	non	basic degree	second degree	third degree	university-entrance diploma
non	1	2	3	6	7
dual system	4	4	5	5	5
vocational school	4	4	5	5	5
vocational college	0	5	5	8	8
university for applied sciences	0	0	9	9	9
university	0	0	0	10	10

Table 9 Hoffmeyer-Zlotnik/Warner Matrix of Education – for *Denmark*

vocational education	general education			
	non	basic degree	second degree	university-entrance diploma
non	1	3	3	7
school/workplace	4	5	5	5
vocational school	4	5	5	5
vocational college	0	5	5	8
university for applied sciences	0	9	9	9
university	0	0	0	10

Table 10 Hoffmeyer-Zlotnik/Warner Matrix of Education – for *Luxembourg*

vocational education	general education			
	non	basic degree	second degree	university-entrance diploma
non	1	2	3	7
school/workplace	4	4	5	5
vocational school	4	4	5	5
vocational college	0	5	5	8
university for applied sciences	0	0	9	9
university	0	0	0	10

Tables 8 to 10 show the matrix for Germany, Denmark and Luxembourg. Common to the three countries are the dimensions of the matrix ranking the school leaving certificates: general graduation by vocational education diplomas. Only not existing and not applicable categories are removed. In Denmark, pupils obtain the basic degree after the 10^{th} grade. In

Luxembourg, the distance between basics and university-entrance diploma is bigger than in Denmark. The German educational system knows two general school qualification levels between the basic degree and the university-entrance diploma.

Missing national certificates lead to missing codes on the 10 categories scale. But the not existing codes emphasize the singularity and individuality of the national education scheme. Some school systems (e.g. the German structure) offer a great number of combinations with different prestige to gain; some national arrangements offer fewer patterns in combining general and vocational certificates.

The Danish matrix still illustrates the need of a two step survey instrument: the question for general education level obtained and the question about the vocational graduation. The ESS questionnaire, fielded in Denmark, groups the answer categories closely to the ISCED 1997 classification. A more detailed survey instrument separating out the general and professional dimension of education may produce a finer defecated measurement of the attained school leaving grades.

5.2 The validity of the Hoffmeyer-Zlotnik/Warner Matrix of Education

The new measurement of education based on the 10 categories matrix is highly correlated with ISCED 1997 classification and the measurement based on "years of schooling". Table 11 also gives the correlation between the occupational prestige (SIOPS) and the household total net income (household-income). For the correlation of ISCED 1997 and SIOPS we have to consider that the skill levels of International Standard Classification of Occupation (ISCO 88) incorporates the ISCED measurement. Therefore we use the empirical prestige scores of an occupation from the survey data and not the theoretical possible value to which a school carrier may end (see Table 11).

Only in Germany, we find a relation between household income and the respondent's educational attainment. In Luxembourg and Germany we detect a strong relation between occupational prestige and our matrix measurement; in Denmark we achieve a lower correlation, but still visible. Comparing the educational measurements, in Germany and Luxembourg our matrix measurement of education is stronger correlated with prestige than the alternative scales. In Denmark, the correlation of our proposal is slightly lower than the years of schooling or ISCED 1997. This may change by using two questions: one about general education and the second about the vocational education. Having the answers on both questions, it is easy to construct the Hoffmeyer-Zlotnik/Warner Matrix of Education by ranking the answer categories. The codes inside the matrix are common across the observed countries and using the prestige score of each combination the national certificates can be reclassified. This limits the researcher's freedom of interpretation of national degrees.

Table 11 Validity of Hoffmeyer-Zlotnik/Warner Matrix of Education: Correlations

	Germany			
	HZ/W	**years**	**ISCED**	**SIOPS**
Years of education	.77			
ISCED	.83	.70		
SIOPS*)	.64	.54	.54	
Household income	.35	.35	.35	.33
	Denmark			
	HZ/W	**years**	**ISCED**	**SIOPS**
Years of education	.71			
ISCED	.93	.77		
SIOPS*)	.49	.50	.53	
Household income	.06	.08	.06	.08
	Luxembourg			
	HZ/W	**years**	**ISCED**	**SIOPS**
Years of education	.74			
ISCED	.93	.75		
SIOPS*)	.61	.56	.58	
Household income	.06	.09	.08	.05

*) SIOPS= Standard International Occupational Prestige Scale by D.J. Treiman
Source: ESS, round 1, computation by the authors

6 Conclusion

Does cross national, cross cultural comparative social research need a new measurement of highest level of education? Looking on the usually applied instruments we found:

"Years of schooling" is an adequate measure when survey researcher and interview respondent have "grades" in mind at the same time. In comparative surveys the question wording must be highly standardized and the translation must be carefully monitored to assure that in all observed countries the same fact is measured.

ISCED 1997 is in most modern and western countries a useful scheme to classify school leaving certificates. In countries with complex educational systems, like Germany, the ISCED 1997 categories cover hardly the social situation. Another disadvantage of ISCED 1997 is the risk misclassification, how national diplomas are sorted into the ISCED 1997 codes. Asking the respondent about the ISCED codes increases the interview burden for the respondent.
The CASMIN Educational Classification is based on a two dimensional measurement, like the Hoffmeyer-Zlotnik/Warner Matrix. But it does not control the freedom of interpretation of the data producer and user during the reclassification process.

The Hoffmeyer-Zlotnik Educational Classification is built consequentially on the combination of general and vocational education and uses the average occupational prestige for ranking the degrees. For the rank order of the degrees we recommend to have national experts involved with a strong knowledge about the national labor market entrance chances.

The Hoffmeyer-Zlotnik/Warner Matrix of Education requires a two step questionnaire, asking for general education followed by a question on vocational education. The table "general" by "vocational" establishes the matrix of educational codes and decreases the risk of misclassification into comparative standard codes by the interviewer and/or the data input, as long as the researcher is guided by the answers given to both questions.

Table 11 shows high correlations between the newly proposed matrix and the ISCED 1997 classification over all countries. Even for Germany, we observe this strong link. This observation confirms the easy use and the low risk of misclassification of our matrix.

A strong relationship between the Hoffmeyer-Zlotnik/Warner Matrix of Education with "years of schooling" is present in all countries. This linkage between the matrix and years of schooling" exists also in countries where "grades" are surveyed; and the relation is higher than the connection between the matrix and ISCED 1997.

Finally, total household net income is independent from all used education scales and from occupational prestige measured by SIOPS.

References

Braun, M., and W. Müller. 1997. "Measurement of Education in Comparative Research." *Comparative Social Research* 16: 163-201.

Brauns, H., S. Schere, and S. Steinmann. 2003. "The CASMIN Educational Classification in International Comparative Research." Pp. 221-244 in *Advances in Cross-National Comparison. A European Working Book for Demographic and Socio-Economic Variables,* edited by Hoffmeyer-Zlotnik, J. H. P., and C. Wolf. New York: Kluwer Academic/Plenum Publishers.

European Commission. 2002. *Key data on education in the European Union – 2002 Luxembourg: Office for Official Publications of the European Communities.* http://www.eurydice.org/Documents/cc/2002/en/CC2002_EN_home_page.pdf

European Social Survey (ESS). http://www.europeansocialsurvey.org/

Eurydice. http://www.eurydice.org/

Eurydice. 2004. *National summary sheets on education systems in Europe and ongoing reforms.* http://www.eurydice.org/Documents/Fiches_nationales/en/frameset_EN.html

Eurydice. 2005. *The education system in Denmark 2000/2001.* http://www.eurydice.org/ Eurybase/Application/frameset.asp?country=DK&language=EN

Eurydice. 2005. *The education system in Germany 2002/2003.* http://www.eurydice.org/ Eurybase/Application/frameset.asp?country=DE&language=EN

Eurydice. 2005. *The education system in Luxembourg 2001/2002.* http://www.eurydice.org/ Eurybase/Application/frameset.asp?country=LU&language=EN

Ganzeboom, H. B. G., and D. J. Treiman. 2003. "Three Internationally Standardised Measures for Comparative Research on Occupational Status." Pp. 159-193 in *Advances in Cross-National Comparison. A European Working Book for Demographic and Socio-Economic Variables*, edited by Hoffmeyer-Zlotnik, J. H. P., and C. Wolf. New York: Kluwer Academic/Plenum Publishers.

Hoffmeyer-Zlotnik, J. H. P. 2003. "The Classification of Education as a Sociological Background Characteristic." Pp. 245-256 in *Advances in Cross-National Comparison. A European Working Book for Demographic and Socio-Economic Variables,* edited by Hoffmeyer-Zlotnik, J. H. P., and C. Wolf. New York: Kluwer Academic/Plenum Publishers.

Hoffmeyer-Zlotnik, J. H. P., and C. Wolf (Eds.). 2003. *Advances in Cross-National Comparison. A European Working Book for Demographic and Socio-Economic Variables.* New York: Kluwer Academic/Plenum Publishers.

International Labour Office. 1990. *International Standard Classification of Occupations (ISCO 88).* Geneva: ILO.

International Social Survey Programme (ISSP). http://www.issp.org/homepage.htm;

NORC and Roper. 1996. *General Social Survey 1972-1996. Cumulative Codebook.* Chicago, IL and Storrs, CT: NORC and Roper.

OECD. 2004. *Education at a Glance. OECD Indicators 2004.* Paris: OECD Publications.

Treiman, D. J. 1977. *Occupational Prestige in Comparative Perspective.* New York: Academic Press.

UNESCO. 1997. "International Standard Classification of Education, ISCED 1997." Pp. 195-220 in *Advances in Cross-National Comparison. A European Working Book for Demographic and Socio-Economic Variables.* Hoffmeyer-Zlotnik, J. H. P., and C. Wolf. 2003. New York: Kluwer Academic/Plenum Publishers. see also: http://www.uis.unesco.org/TEMPLATE/pdf/isced/ISCED_A.pdf

ZUMA/GFM-GETAS. 1996. *Standard-Statistik Sozialwissenschaften-Bus.* Hamburg: GFM-GETAS/WBA.

ON THE COST OF BEING CRUDE: A COMPARISON OF DETAILED AND COARSE OCCUPATIONAL CODING IN THE ISSP 1987 DATA

HARRY B.G. GANZEBOOM [1]

1 Introduction

Occupational categories constitute the backbone of sociological research in social stratification. However one conceptualizes occupational status (prestige, class, socio-economic status, cf. Ganzeboom & Treiman, 2003), the way in which occupations are initially classified is always a core ingredient of the measurement procedure. Typically, in high quality surveys, information on occupations is recorded in a sequence of open-ended questions. These questions will ask for job title, main duties and activities, employment status and supervising status. Part of this information (usually job title and main activities/duties) is then converted to a detailed occupational classification in post-processing the information. The detailed occupational classifications used are mostly provided by national or international statistical agencies and often distinguish between 500 and 1500 different occupational categories. Coding these categories from the verbatim information is a time-intensive operation that consumes a substantial part of survey budgets.

The basic question of the research reported here is whether this coding operation is worth the trouble. How much do we gain from coding occupations in a detailed classification as opposed to more easy to operate crude procedures? The assumption, of course, that under-

1 Earlier versions of this paper were first presented to the ISA Research Committee on Social Stratification, Warzaw 1999, and last to the Zentrum für Umfragen, Methoden und Analysen, June 2005. The initial idea for this paper arose in a discussion with Jonathan Kelley and Mariah Evans in Prague in December 1989, in which we let the Velvet Revolution go by and became more fascinated by the ”Kelley & Mariah Coding Disaster”. I thank Kelley and Mariah, and many discussants at conferences for their critical comments.

lies the use of detailed occupational classifications, is that there are sociologically relevant properties of occupations (say educational requirements, earning potentials) that vary among occupations (mostly) at the detailed level. It this is so, using cruder classifications would introduce 'aggregation' bias by obscuring part of that variation and this would result in attenuated associations between occupational variables and their causes and consequences. In the analyses reported below we estimate the degree of attenuation using a standard model of status attainment, in which two occupations occur: father's occupation and respondent's current/last occupation.

Some prior experiences have led me to suspect that the degree of attenuation may not be large and may not warrant the costs involved in the implementation of detailed occupational classifications. In an analysis of intergenerational class mobility (Ganzeboom, Luijkx & Treiman, 1989), we used the degree of detail of the underlying occupational codes as a control variable and found that – all else being equal – the association between father's occupation and son's occupation tends to be highest in case of moderate crudeness of the occupational data. In another analysis, that reported on validation of the International Socio-Economic Index (ISEI) of occupational status (Ganzeboom, de Graaf & Treiman, 1992), we estimated formal attenuation coefficients for various categorizations of the data in a three variables status attainment model (father's occupation, son's education, son' occupation) and found only moderate attenuation (around .95) for measures with as few as ten or six categories; only for a three category recoding of the data the attenuation was appreciable, although still only .85[2].

There can be three different reasons why previous research has found such minor attenuation effects. A first possibility is that occupations within broad categories do not vary much among one another in sociologically interesting ways. In these cases crude measurement suffices. An alternative possibility is that occupation coding at a detailed level is more prone to measurement error than crude classification. Respondents, interviewers and coders may have a better understanding of an occupation at a crude level than at a detailed level. A third possibility, somewhat related to this second explanation, is that crude classifications – in particular when used in in-field coding – pick up variance from other variables and thereby results in stronger associations. For instance, crude occupational schemes are likely to present respondents with clues about skill levels, supervisory status and self-employment, and these sociologically meaningful additions are likely to become part of the criteria that lead the coder, interviewer or interviewee to a decision upon the most plausible category in a precoded response format. We feel that it is particularly worrisome

2 The cited attenuation coefficients should be read as the degree to which using a crude measure attenuates covariances/correlations with the occupation variable, relative to a detailed measure.

worrisome when a person's educational *qualifications* become mixed up with his/her job's educational *requirements* (which is likely to be the case when skill levels are presented), since this confounds measurement with substance in one of the central concerns of social stratification research. If this situation would hold, it is not implausible that crude measures lead to stronger associations in empirical data on father-to-son occupational mobility, as observed in Ganzeboom, Luijkx & Treiman (1989). However, in this case one would expect that the extra strength of the effect disappears, once education is controlled.

To investigate these concerns more thoroughly, I compare in the analyses below the structure of status attainment models using detailed classifications and crude classifications. To do so, I take advantage of the fact that there exist a large-scale cross-national dataset that has measured occupations independently in a crude and a detailed way: the International Social Survey Programme 1987 [ISSP87] (ISSP, 1987). While these data are rather old by now, there is no reason to assume that they have become irrelevant to the issues at stake here. To my knowledge, the issue of crude versus detailed occupation coding has not been addressing using this dataset.

Table 1 Type of Work Question – ISSP 1987

Here is a list of different types of jobs. Which type did your father have when you were 16 years / [did you have in] the first job you had after you finished your full-time education / [do you have]] in your job now?

01. Professional and technical (for example: doctor, teacher, engineer, artist, accountant)
02. Higher administrator (for example: banker, executive in big business, high government official, union official)
03. Clerical (for example: secretary, clerk, office manager, civil servant, bookkeeper)
04. Sales (for example: sales manager, shop owner, shop assistant, insurance agent, buyer)
05. Service (for example: restaurant owner, police officer, waiter, barber, caretaker)
06. Skilled worker (for example: foreman, motor mechanic, printer, tool and die maker, electrician)
07. Semi-skilled worker (for example: bricklayer, bus driver, tannery worker, carpenter, sheet metal worker, baker)
08. Unskilled worker (for example: labourer, porter, unskilled factory worker)
09. Farm (for example: farmer, farm labourer, tractor driver)

Was your father / were you / are you self-employed, or did he /did you / do you work for someone else?

1. Self-employed, own business or farm
2. Work[ed] for someone else

2 Data and Variables: The International Social Survey Programme 1987

The 1987 issue of the International Social Survey Programme (ISSP, 1987), with a module on Social Inequality, included an experiment with a precoded question on occupations, of which details are reproduced here in Table 1. The respondent was asked to choose an appropriate category out of nine, each of which is prompted by a general label and a variable number of typical occupations in the category. In a limited set of countries, this crude question was then followed up with an open question, and the results of these questions were coded in standard occupational classification tools. This crude question was asked for respondent's occupation, father's occupation and respondent's first occupation in all countries, except Hungary[3], but the detailed question in only five countries, and only for father's and respondent's current/last occupation. Since I want to make a comparison between detailed and crude measurement procedure, the analysis will be restricted to the combination of countries and variables for which information was collected independently in both modes. Since there is no detailed information on first occupation in any of the countries, the analyses will concentrate on father's and respondent's current/last occupation. There are five countries in the ISSP87 for which the two sets of information on these two variables were collected: Australia, the USA, Austria, Germany, and Switzerland. In the latter three countries, the detailed occupation codes are provided is the International Standard Classification 1968 [ISCO68], while the Australians and Americans have used national detailed occupations classification (CPS70 and ASCO86, respectively). These two were converted in ISCO68, using previously developed recoding schemes (see ISMF, 2005).

In all countries an additional question was asked on self-employment of the respondent and his/her father and I decided to combine this information with the crude occupational categories. This is particularly important in the case of professionals, higher administrators, sales workers, and farm workers, since it makes it possible to distinguish self-employed professionals, large business owners, shop owners and farm owners from salaried professionals, managers, sales clerks and agricultural laborers. The two questions in the ISSP87 module thus combine into 2*9=18 separate categories that in practice reduce to 13 categories. Each of these 13 categories was scaled into the International Socio-Economic Index of occupational status ISEI (Ganzeboom, de Graaf & Treiman, 1992),

3 At first impression from the data documentation, it appears as if a similar but different crude question was asked in Hungary, but this turns out not to be the case: the respective variables in Hungary do not contain independent information, but are straight recodes from the questions on detailed occupations.

using ISCO68 as a conversion tool (ISMF, 2005). Similarly, the detailed ISCO68 occupation codes were scaled into ISEI. On average, the two parallel ISEI measures correlate between .65 and .75: the correlations are a bit higher for father's occupation than for respondent's occupation. The resulting variables are labeled FISEI and ISEI (father's and respondent's occupation derived from detailed measures), and FASEI and ASEI (derived from the alternative crude measurement), respectively.

The ISSP87 has not only experimented with different procedures for the measurement of occupation, a somewhat related procedure was used for education, that was also measured with two parallel questions, one about the highest grade attained and one about number of years completed. The interpretation of this operation is a bit different than for occupation, since although the highest grade attained question usually implies a less detailed measurement, at the same time it taps distinctions that are locally important. In particular when educational systems are divided in vocational and academic tracks (which is the case in Austria, Germany, and Switzerland), the highest grade attained will tap different and probably more relevant information than the years-of-school completed measure. For the analysis here, the years-of-school measure completed was maintained in its original format. The highest grade completed was rendered in a comparable metric by ranking the different grades according to the years of school completed and express the categories in percentile score. These variables enter the analyses as EDYRS (years of school completed) and EDRANK. On average these are correlated around .80, which implies that they indeed tap somewhat different aspects of the education.

The third status attainment variable is personal earnings, which was originally measured using local currencies and with slightly different prompts. The number of categories varies between 12 and 25. The measures were made cross-nationally comparable by expressing the categories in percentile scores within countries. The resulting variable is labeled INCRANK.

Finally, AGE and sex (FEMALE: women=1, men=0) are used as control variables. The effective samples were restricted to be between 21 and 64 years of age and the data on men and women were pooled on the argument that difference between men and women in distributions can be adequately modeled by using sex as a control variable.

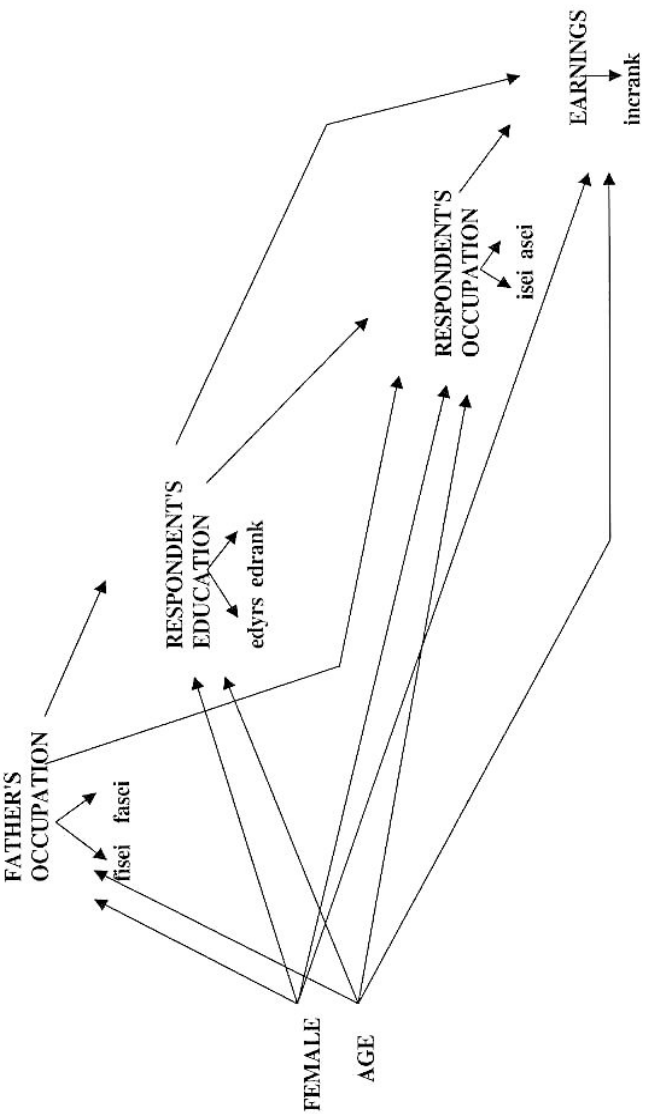


Figure 1: The elementary status attainment model

For each of the five countries a nine-variable correlation matrix was derived using pairwise deletion of missing data[4]. In total there are some 5000 cases (as pairwise deletion of missing values has been applied, this varies between relationships) in the analyses. These correlation matrices were analyzed via a structural equation model, estimated in LISREL8 (Jöreskog & Sörbom, 1993), using maximum likelihood fit procedures. This fitted model is displayed in Figure 1. It is almost fully saturated at the structural level, except that I do not assume any direct effect of father's occupation on respondent's income, as well as no association between FEMALE on the one hand and age and father's occupation on the other. The model is estimated in the following versions:

I. A single indicator model, with detailed occupations as measures and the 'best' single indicator for education.

II. A single indicator model, with crude occupations as measured and the 'best' single indicator for education.

III. A model with latent variables for the two occupations and education, with both the detailed and the crude indicator as measures.

Comparison of I and II leads to an assessment whether and to what extent detailed or crude occupation measures lead to higher associations. Given that these are single indicator models, we can use R-squared measures to make the comparison. The comparison is more direct in model III, where the measurement relationships (Lambda's in LISREL) can directly be interpreted as attenuation coefficients, not only relative to one another, but also relative to a true-score model corrected for measurement error. In addition, model III gives an estimate to which extent multiple indicators measurement improves the estimates.

4 The correlation matrices are available from the author's website: http://home.fsw.vu.nl/hbg.ganzeboom, from which a full version of this paper, including numerical appendices can be downloaded.

Table 2 Fit Statistics of Lisrel Models

	Same pattern		Invariant	
I. Single indicator models, detailed measures				
a. Saturated structural model	10	17	86	277
b. Model (a) – be(6,3)	15	36	87	278
c. Model (b) – be(6,4)	20	136	88	370
II. Single indicator models, crude measures				
a. Saturated structural model	10	17	86	344
b. Model (a) – be(6,3)	15	26	87	345
c. Model (b) – be(6,4)	20	116	88	418
III. Multiple indicator models				
a. Saturated structural model	85	390	197	1560
b. Model (a) – be(6,3)	90	406	198	1567
c. Model (b) – be(6,4)	95	435	199	1577

Each of the models is estimated for the five countries separately, as well as for all the countries pooled (denoted as XNAT), using the 'invariant' option in Lisrel's multiple group specification. This pooled solution provides a parsimonious insight in the average results, in particular when the between-country differences are not spectacular, as well a useful benchmark for the country-wise results. The fit statistics are provided in Table 2. The modelling strategy has been that we compare the almost saturated model (a) with a model (b) in which there is no direct effect of father's occupation in earnings. In model (c) I remove in addition the effect of education on earnings – which leads to an appreciable loss of fit for the single indicator models, but not so much for the multiple indicator models. While none of the estimated models fits the empirical correlation matrices by standard statistical standards, one should take in to account that the analyses deals with more than 5000 cases.

Table 3 Standardized Estimates of an Elementary Status Attainment Model with *Singular* Indicators for Education and Occupation

A. Structural Coefficients Using Detailed Measures

	AUS	GER	USA	AUT	SWI	XNAT
EDUC (R2)	(.249)	(.218)	(.194)	(.234)	(.197)	(.205)
– FEMALE	-.083	-.057	-.052	-.074	-.133	-.083
– AGE	-.222	-.177	-.131	-.180	-.006	-.142
– FISEI	.335	.408	.386	.413	.424	.402
ISEI (R2)	(.394)	(.414)	(.334)	(.395)	(.328)	(.360)
- FEMALE	.009	.041	.068	.018	.006	.029
- AGE	.203	.138	.148	.100	.099	.136
- FISEI	.095	.156	.128	.251	.117	.149
- EDUC	.570	.578	.532	.496	.511	.537
EARNINGS (R2)	(.464)	(.425)	(.321)	(.362)	(.419)	(.365)
- FEMALE	-.468	-.483	-.355	-.441	-.421	-.435
- AGE	-.034	.240	.233	.026	.189	.143
- EDUC	.207	.054	.157	.159	.094	.138
- ISEI	.194	.341	.293	.282	.349	.290

B. Structural Coefficients Using Crude Measures

	AUS	GER	USA	AUT	SWI	XNAT
EDUC (R2)	(.184)	(.195)	(.216)	(.205)	(.143)	(.178)
- FEMALE	-.083	-.035	-.056	-.050	-.109	-.070
- AGE	-.222	-.189	-.124	-.188	.043	-.136
- FASEI	.335	.378	.415	.378	.366	.373
ASEI (R2)	(.387)	(.334)	(.373)	(.438)	(.344)	(.363)
- FEMALE	.077	-.090	.138	.002	-.063	.011
- AGE	.172	.118	.130	.047	.063	.106
- FASEI	.158	.196	.115	.232	.136	.167
- EDUC	.571	.476	.562	.543	.512	.531
EARNINGS (R2)	(.382)	(.420)	(.293)	(.341)	(.422)	(.359)
- FEMALE	-.486	-.442	-.365	-.432	-.404	-.433
- AGE	.030	.232	.246	.042	.190	.146
- EDUC	.165	.025	.195	.170	.075	.122
- ASEI	.254	.367	.218	.251	.366	.297

AUS: Australia, GER: Germany (West), USA: United States, AUT: Austria, SWI: Switzerland, XNAT: Cross-national (pooled).

3 Results

3.1 Status attainment models

Table 3A gives the structural coefficients for the model b, in which only the detailed occupation codes are used. Table 3B gives estimates for the same model for the crude occupation measures. The models consist of three separate equations that show a familiar pattern to the experienced stratification researcher. The first equation relates education to father's occupation (and sex and age); it suggests that detailed coding is clearly superior to crude coding. On average the effect of father's occupation is attenuated by a factor .91 (.364/.402). There is also a 4% additional explained variance when one uses the detailed codes to scale father's occupation.

The second equation, for respondent's occupation, shows much less spectacular differences between models. The amount of variance explained is almost the same (.360 versus .359) and there is hardly any difference in the estimated coefficients. Note, however, that the direct effect of father's occupation on respondent's occupation is larger for the crude codes than for the detailed codes. This suggests that the degree of attenuation for father's occupation, as estimated from the first equation, does not apply to the relation between father's occupation and son's occupation. It also suggests that crude measures are slightly more prone to lead the respondent to bias the report on father's occupation towards his/her own occupation.

The coefficients of the third equation, on respondent's earnings, are even more similar between the two coding modes, both with respect to variance explained and the size of the coefficients. On average, the numbers are again slightly in favor of the detailed measure, but this is not the case in all separate countries and the differences are very small. According to both models, earnings are distinctively lower among women and young people (note that there is no control for hours worked in the model) and they are positively affected by both occupation and education. It is of some importance to focus a bit on the net effect of education in income: this effect implies that the higher educated make more money than lower educated *within* jobs of the same level. This net effect of education is routinely observed in income models and may be given different explanations. While it may be true that higher educated are higher remunerated for the same work, because they perform better or because income is awarded for formal credentials, the effect may also occur because of bad measurement of occupation. Using a multiple indicator approach, we will be able to test this latter explanation.

Table 4 Standardized Estimates of an Elementary Status Attainment Model with Multiple Indicators for Education and Occupation

	AUS	GER	USA	AUT	SWI	XNAT
I. Measurement models						
FISEI						
--> detailed	.886	.886	.886	.886	.886	.885
--> crude	.826	.798	.911	.790	.858	.835
EDUC						
--> years	.907	.907	.907	.907	.907	.901
--> rank	.819	.914	.883	.989	.905	.899
ISEI						
--> detailed	.836	.836	.836	.836	.836	.835
--> crude	.872	.737	.862	.863	.822	.829
II. Structural model						
EDUC (R2)	(.246)	(.341)	(.282)	(.398)	(.308)	(.298)
- FEMALE	-.082	-.052	-.057	-.072	-.136	-.090
- AGE	-.244	-.183	-.084	-.192	.041	-.124
- FISEI	.469	.505	.542	.458	.534	.503
ISEI (R2)	(.551)	(.631)	(.611)	(.624)	(.589)	(.591)
- FEMALE	.042	.006	.127	.025	-.005	.038
- AGE	.220	.207	.171	.152	.094	.165
- FISEI	.196	.091	.122	.161	.056	.136
- EDUC	.554	.835	.664	.846	.766	.710
EARNINGS (R2)	(.413)	(.457)	(.354)	(.368)	(.481)	(.404)
- FEMALE	-.491	-.469	-.393	-.441	-.406	-.439
- AGE	-.023	.229	.205	.013	.175	.120
- EDUC	0	0	0	0	0	0
- ISEI	.434	.395	.465	.390	.478	.434

AUS: Australia, GER: Germany (West), USA: United States, AUT: Austria, SWI: Switzerland, XNAT: Cross-national (pooled).

Table 4 shows the same status attainment model, but now estimated with a multiple indicator design. By comparing lambda's we can estimate the degree of attenuation directly, relative to the true score. These are spelled out in the measurement part of the model. The model has the advantage of pooling all the evidence into one estimate. The degree of attenuation for crude measurement of father's occupation relative to detailed measurement is found to be .95 (.834/.886), but for respondent it is a meager .99 (.827/.833). This pattern varies a bit between the countries, and in a few instances the estimates even suggest that crude codes are to be preferred over detailed codes. But however one looks at these numbers, the differences between the two modes are very minor. However, the same coefficients can now be compared to unity (1.0), which represents the attenuation relative to the true score. Averaging over coefficients, we can conclude that the attenuation relative to the true score is not so small, but amount to at least 15%. I.e., for each and every correlate of an occupational status, we find at best 85% of the true correlation, if we use either one on the two measures!

Note in passing that the estimated lambdas for education are much closer to unity than those for occupation and almost in balance for three countries. The Australian estimates suggest that years completed is to be preferred over ranked grades, whereas the Austrian case suggests the reverse.

While the analysis of the measurement relationship in Table 4 confirms our conclusions from comparing the single indicator models for the attenuation of crude measures relative to detailed measures (and actually suggest that attenuation is even less spectacular than what these models imply), the spectacular part of the table is part B, on the structural model, as it show how the attenuation relative to the true score affects findings. Crude *or* detailed occupations hardly make a difference, but using *both* does! The effect of father's occupation on education increases by 1.32, the effect of education on occupation by 1.35 and the effect of occupation on income by 1.65. Parallel increases are found for variance explained. Note in particular that in the income equation the effect of education is now estimated to be slightly negative (-.072). Since a negative value is theoretically implausible, I have re-estimated the model without the education effect, which reduced the occupation effect to .434. This still implied a disattenuation relative to the single indicator model of 1.43.

Note also that in the second equation the direct effect of father's occupation on respondent's occupation, relative to the single indicator models, has dropped from .16 to .13. This is an illustration that unreliable measurement attenuates indirect effects more strongly than direct effects. Once a proper measurement model is taken into account, the upward bias in the direct effect disappears. Since the lambda's of the parallel education

measures are fairly high, the shift for this is not spectacular, but still in the predicted direction.

All these results are due to the attenuation of the measures relative to the error-corrected true score model, which suggest that for both crude and detailed measurement about 15% attenuation occurs.

3.2 Intergenerational mobility tables

The analyses reported above compare the effect of different modes of coding assuming that occupational status is adequately reflected in a (semi-)continuous measure. Much research on occupation, however, uses categorical measures of occupational status. The main argument for preferring categorical representation is the belief that occupational differences result inherently from discrete and multidimensional processes that are best represented by (class) typologies, in which multiple occupational variables are combined. There is in fact ample empirical evidence in favor of this position. In particular in occupational mobility research (in which transitions between two or more occupations are investigated), it has often been shown that the association between variables has properties than cannot be represented by correlation and regression coefficients: multidimensionality and asymmetry. For these types of analyses, various loglinear models have been proposed and used (Hout, 1983).

How adequate are crude occupational classifications in generating discrete occupational measures? It is to be noted beforehand that detailed occupational classifications have the advantage that they can combined in many different typologies, while crude classification leave only few degrees-of-freedom in this respect. However, this is not the issue in this paragraph. Here we assess whether crude and detailed classifications behave differently, when used in constructing a class typology that is consistent with the crude codes.

In order to make comparisons, I combined both detailed and crude occupation with the self-employment code to derive an EGP-class typology with seven categories[5]. By comparing the results from the detailed and crude approach in measurement, we can learn what effects the choice of a crude/detailed coding system has on the results of mobility analyses.

5 The counts for these mobility tables can also be obtained from the full paper on the website of the author: http://www.fsw.vu.nl/~ganzeboom.

Table 5 Fit Statistics and Estimated Association Coefficients for Intergenerational Occupational Mobility, Using Detailed and Crude Coding Procedures

I. Fit statistics (LR2)

		ndf	AUT	GER	USA	AUT	SWI	NAT
a.	**Independence** (O*T,D*T)	72	374	380	300	506	245	1809
b.	**Common association** (a+O*D)	36	27	26	29	27	32	379
c.	**Two common components** (a+DIA+U+INH)	64	76	74	91	80	70	423
d.	**Two components** (c+(U+INH)*T)	62	71	71	89	79	71	418

II. Parameters model I.d

U	.065	.123	.065	.106	.067	.097
U*T	-.005	.005	-.004	-.010	-.008	-.004
INH	dia	dia	dia	dia	dia	dia
INH*T	-.089	-.118	-.078	.048	.009	-.049

Notes:
O: Origin. D: Destination. T: Type of coding: (1) detailed (0) crude. IMM: class immobility coefficients. U: uniform association coefficient. INH: uniform inheritance coefficient. Dia: class-specific inheritance coefficients. Ndf Degrees-of-freedom. Note that the comparisons are not on independent samples and that the entries only have descriptive value.

AUS: Australia, GER: Germany (West), USA: United States, AUT: Austria, SWI: Switzerland, XNAT: Cross-national (pooled).

Table 5 gives fit statistics and selected parameter estimates for four relevant loglinear models, which compare the tables derived from the two coding modes. Note that the fit statistics here only have descriptive value, since we are not comparing independent samples. Nevertheless, it is immediately apparent from model Ib that notwithstanding their dissimilar marginal distribution the association pattern in the intergenerational mobility table are strikingly similar, as they all approach the number of degrees of freedom. However, the Common Association Model Ib is not a very sharp tool in deciding about differences between tables, since it consumes many degrees-of-freedom to make the comparison. Model Ic uses the uniform association model, extended with generic immobility

coefficients to set up a two-degrees-of-freedom comparison (the approach is the same as in Ganzeboom, Luijkx & Treiman, 1989). Model Id conditions the two principal association coefficients, labeled U and IMM, by type of coding. The comparison shows again that there is very little difference between the two ways of recording occupations. This is confirmed again in panel II of the table, where estimates for the parameters in model Id are given. It turns out that these are not only nominally insignificant, but also that the differences between the tables do not amount to much more than a few percentage points. The estimates in panel II also show that there is not a uniform pattern in the association coefficient: in about half the cases the coefficients are in the direction of stronger association for the crude codes, for the other half it is the other way around. This lack of pattern is confirmed by the very small difference in the model for the pooled data that aggregates over these variations among countries. However, is there is any difference, it is that the association is slightly less strong in the data derived from the detailed classification.

These results imply again that for some purposes it makes very little difference whether one starts out from detailed or crude codes. Apart from the differences in marginal distributions, there seems to be even less differences in effect of coding than detected by the procedures that conceptualized occupational status as continuous measures.

4 Conclusions and Discussion

The conclusions from this analysis of the comparison of crude and detailed occupation codes can be formulated as follows:

1. There is very little difference in unreliability (random measurement error) between detailed and crude occupation codes. On average the results from the status attainment models favor the detailed codes by a small margin, but this margin is indeed so small that it seems hard to argue that establishing detailed occupation codes is warranted for this purpose.

2. Crude occupation codes seem to be slightly biased towards immobility, i.e. there appears to be some tendency to put father in the same class as one self. This tendency, again, is very slight.

3. The attenuation effect of single indicator measurement relative to true scores in a multiple indicator measurement is rather dramatic (between .80 and .85) and measuring crude and detailed codes at the same time seems to be a natural way to create a multiple indicator design.

Why is it the case that crude and detailed measurement procedures make so little difference? Can it be the case that a true occupational status score becomes more corrupted, when a less detailed measurement instrument is used? I propose that at least two processes are at work here.

First, deriving detailed occupation codes is a much more complicated procedure than deriving crude codes. Crude codes are basically a self-evaluation by the respondent, who understands best what s/he (or her/his father) is/was doing for a living. Detailed occupation measures, by contrast, require understanding on the part of the interviewer, who records the information, as well as the coder. These two steps of communication take their toll, as any communication leads to misunderstanding. The attenuating effect of these procedures would be testable in a repeated measurement design, in which the same information is recorded independently by different interviewers and coded independently. Unfortunately, such data are not available at this time.

Second, it may be that standard occupational classifications by themselves are less adequate classification tools than their level of detail suggests. It seems plausible that many aspects of occupation determine their educational requirements, their earning potential and their use as resource in social mobility. However, it remains to be seen that those aspects are well covered by the distinctions often made in occupational classification. Previous experiences suggest that there is very little systematic variance with respect to educational requirements and earnings potential in the last two, or even last three digits of ISCO, and it seems likely that the same applies to national classifications. I.e., essentially the same results would arise, if coders / interviewers would have restricted themselves to coding only one or two digits. This does not imply that there is no variation among occupational position in this respect, but only that these are not picked up by the standard classifications.

The way to pick up the importance of the various aspects that are important for the status attainment attributes of someone's occupation is therefore the multiple indicator design, much as we are used to apply it in attitude research. One practical problem with multiple measurements of structural and demographic characteristics may be that good parallel questions are hard to construct. While it is trivial to ask repeated questions on someone's attitude towards abortion, this would be irritating for occupation, education, etc. It strikes me that asking both crude and detailed question in ISSP87 is in fact a natural and acceptable way to circumvent to difficulty. I am currently in the process of collecting and analyzing such data in a national context.

References

Ganzeboom, H. B. G., P. de Graaf, and D. J. Treiman (with J. de Leeuw). 1992. "A Standard International Socio-Economic Index of Occupational Status." *Social Science Research* 21(1):1-56.

Ganzeboom, H. B. G., R. Luijkx, and D. J. Treiman. 1989. "Intergenerational Class Mobility in Comparative Perspective." *Research in Social Stratification and Mobility* 8:3-79.

Ganzeboom, H. B. G., and D. J. Treiman. 2003. "Three Internationally Standardised Measures for Comparative Research on Occupational Status." Pp. 159-193 in *Advances in Cross-National Comparison. A European Working Book for Demographic and Socio-Economic Variables,* edited by Hoffmeyer-Zlotnik, J. H. P., and C. Wolf. New York: Kluwer Academic Press.

Hout, M. 1983. *Mobility Tables.* Beverley Hills: Sage [University Paper #31].

ISSP [International Social Survey Programme] 1987. *International Social Survey Programme 1987: Social Inequality* [machine-readable dataset]. Cologne, Germany: Zentralarchiv für Empirische Sozialforschung. University of Cologne [producer]. Ann Arbor, MI.: Inter-university Consortium for Political and Social Research [distributor].

ISMF [Ganzeboom, H. B. G., and D. J. Treiman]. 2005. *International Stratification and Mobility File: Conversion Tools.* Retrieved August, 2005, from http://www.fsw.vu.nl/~h.ganzeboom/ismf

Jöreskog, K. G., and D. Sörbom. 1993. *Lisrel 8 User's Reference Guide.* Chicago: Scientific Software International.

ETHNICITY AND THE COMPARATIVE ANALYSIS OF CONTEMPORARY SURVEY DATA

*PAUL S. LAMBERT**

1 Introduction

Concepts of minority ethnic group are important to a great many social science analyses. Ethnic differences are often a focus of analysis; even when they are not, a case can very often be made that interactions between ethnicity and the topic of study will be non-ignorable (Anthias, 2001). Yet developments in the sociological conceptualisation of ethnicity generate serious problems for cross-national survey researchers. Typically, sociological discussions have argued the need for ever finer qualitative differentiations between different ethnic situations in any particular country. The task of mapping and analysing the *minutiae* of such ethnic differences, in terms of comparable variables in cross-national survey datasets, becomes highly problematic.

This challenge is not, however, one which social researchers should shy away from. The internationally comparative analysis of ethnicity effects has considerable substantive appeal, as different outcomes might be related to variations in national contexts and policies (e.g. Castles, 1993). At a more practical level, history suggests that comparative analyses of ethnicity, using whatever survey data categories are available, will continue apace – regardless of any misgivings about the validity of the variable indicators available.

* The data analysed in this paper was gratefully provided by: European Social Survey; UK Data Archive; Economic and Social Data Service (UK); Luxembourg Income Study; International Public Use Microdata Service. Any errors in the subsequent use of these datasets are the author's responsibility. The research was assisted by an IRISS grant from the CEPS research institute.

2 Concepts of Ethnicity in Social Survey Research

Relatively few methodological writings reflect upon measures of ethnicity in cross-national surveys, and, when attention is paid, it is often restricted to an overwhelmingly critical perspective (Allen & Macey, 1990; Rea et al., 1999; Favell, 2003: 25-34). There are, however, a few prescriptive evaluations of the comparative analysis of ethnicity through micro-social surveys (e.g. Lambert & Penn, 2001; Hoffmeyer-Zlotnik, 2003; Bonifazi & Strozza, 2003). There are also several outputs based upon cross-national survey research (e.g. Stille, 1999; Heckmann et al., 2001; Evans & Need, 2002; SYSDEM, 2003; Jacobs & Tillie, 2004; van Tubergen et al., 2004[1]).

In a recent review of the measurement of race and ethnicity in comparative survey research, Hoffmeyer-Zlotnik (2003) highlighted the complications caused by lack of consistency in just which underlying concept different social scientists have in mind when they discuss race and ethnicity (see also Aspinall, 2002). This underling concept of interest can be termed the *'referent'* of the ethnicity measure (cf. Lambert & Penn, 2001). In previous literatures, the referent chosen has ranged over topics such as citizenship, national origins, country of ancestral origins, racialised visibility, language spoken, subjective cultural identity, and religion. In this writing, we use *'ethnicity'* as an umbrella term to refer to differences associated with any of these concepts. This reflects the use of the term in British literature, as a self-assigned category free to incorporate diverse influences (cf. Hutchinson & Smith, 1996; Banton, 1997)[2]. Table 1 describes the range of alternative ethnicity referents which a group of four major cross-national survey datasets collect.

Survey definitions of ethnicity are usually made in terms of categorisations based upon the boundaries of one or more alternative referents. The term categorisation is crucial, as it leads to an apparently irresolvable tension for survey researchers. For the purposes of data analysis, simple, parsimonious categorisations, with large numbers of cases in all groupings, are preferred. Candidate ethnic referents, however, have anything but those properties. They incorporate many categories with very sparse numbers of cases. And, according to social theorists, they should be cross-classified across a large number of alternative permutations.

1 Other ongoing projects in this field include the TIES project coordinated by M. Crul (see http://www.niwi.knaw.nl/en/oi/nod/onderzoeker/PRS1258604/toon); and the "Ethnic minority disadvantage in the labour market" project coordinated by A. Heath (see http://www.britac.ac.uk/events/2003/031102emd-prog.html).

2 This term is not ideal, since in other circumstances, 'ethnicity' is taken to refer specifically to subjective cultural identity.

Table 1 Ethnic Referent Data Availability, by Country and Study

Ethnic referent:

C	**Citizen of which country**	**P**	**Parental country of birth**	**E**	**Ethnic self-identity**
B	**Country of birth – which**	**L**	**Which language used**	**R**	**Religious denomination**
T	**Time in this country**			n	*No relevant data*

Letters indicate presence of data on relevant referent. Lower case letters when categories are dichotomy only – e.g., born in host country or not is 'b' rather than 'B'. Blank cells for non-coverage of country within study

	ESS	ISSP	WVS	LIS		ESS	ISSP	WVS	LIS
Australia		BR	BTLER	BT	*Latvia*		CLR	BTLER	
Austria	CBTPLeR	R		Cb	*Lithu.*			BTLER	
Belgium	CBTPLeR			C	*Luxem.*	CBTPLeR			CT
Bosnia			BTR		*Maced.*			bTR	
Bulgaria		RE	BTR		*New Z.*		ER		
Canada		LER		bT	*Nthlds*	CBTPLeR	R		
Croatia			BTER		*N. Irel.*		ER		
Cyprus		ER			*Norway*	CBTPLeR	R	BTR	Bp
Czech R	CBTPLeR	bR		C	*Poland*	CBTPLeR	BR	bER	n
Denmark	CBTPLeR	BR		CT	*Portugal*	CBTPLeR	cR		
Estonia			BTLER	E	*Russia*		ER	BTLER	BE
Finland	CBTPLeR	LR	bTLR	L	*Serbia+M*			bTLER	
France		R		C	*Slovenia*	CBTPLeR	ER	bTlER	n
Germany	CBTPLeR	CR	RBTE	CB	*Spain*	CBTPLeR	R	BTLER	n
Greece	CBTPLeR				*Sweden*	CBTPLeR	cR	bTLR	CBTp
Hungary	CBTPLeR	ER		E	*Switz*	CBTPLeR	CLR	BTLR	c
Ireland	CBTPLeR	R		CBT	*UK (GB)*	CBTPLeR	ER	E	E
Israel	CBTPLeR	BR		T,R-B	*USA*		ER	BTLER	cE
Italy	CBTPLeR	R		B					

ESS : all studies from 2002.
ISSP : all studies for 2000, except Australia, Cyprus, France, Hungary, Latvia, Poland (1999) and Italy (1998).
WVS : Wave 3 only 1995-7 (other countries are covered by WVS in earlier waves and not W3). In some WVS countries, relevant data is nominally present, but all categories undocumented, so listed here as missing.
LIS : uses latest available LIS study, within range 1994-2001.

Indeed, extended sociological debates on the nature of ethnic differences have generated what may be characterised a 'paradigm of diversity' in the field. In this model, the promotion and discussion of diversity of ethnic locations dominates contemporary discourse. At the least, the role of a great many categorical differences in defining ethnic locations is emphasised (Modood et al., 1997, 2002; the significance of boundaries may also vary over time, e.g. Aspinall, 2002). In stronger interpretations of the paradigm of diversity, many writers have argued for the inherent instability and contextuality of ethnic locations, as they are complicated by evidence of cultural hybridities (Aspinall, 2003), transnationalism (Vertovec, 2003), and/or complex social networks (Sanders, 2002). Such perspectives have lead some to the conclusion that survey variable operationalisations can offer at best a weak analytical device for representing ethnic diversity (e.g. Ahmad, 1999). Neverthe-

less, within many countries, attempts have been made to define an optimal ethnic categorisation in terms of the one or more ethnic referents thought most important[3] – these constitute 'official' ethnic categorisations available on survey outputs.

An immediate problem for cross-national researchers is that the most relevant ethnic diversities in different countries are rarely equivalent, leading to substantial differences in the type of ethnic data which is collected by different national surveys. The cross-national comparability of categorisations which do not use the same ethnic referents is clearly questionable[4]. Indeed, many previous discussions of ethnicity measures for comparative research have ended at this point, abandoning the project (e.g. Rea et al., 1999).

Differences in data availability between countries arise both as a result of alternative perspectives in social science theories[5], and as a consequence of nation specific institutional traditions (Lambert & Penn, 2001; Hoffmeyer-Zlotnik, 2003; Favell, 2003). The latter are often strongly politicised and rigidly enforced – see for example Favell's (2001) description of the constrast between British and French views on the official recording of 'racialised' categories. Indeed, national institutions' influence over the research data available have caused considerable difficulties to previous research programmes – illustrated by attempts at coordinating reviews of labour market situations reported in two recent European Employment Observatory Reviews (Stille, 1999; SYSDEM, 2003).

A common reaction in cross-national research methodology has been attempts to adjudicate between alternative ethnic referents in a manner which is parsimonious, cross-nationally consistent, and politically acceptable. Hoffmeyer-Zlotnik (2003) suggests a strategy for delineating the key dimensions of six ethnic referents in a manner amenable to social survey questions. This leads to the prescription of a series of differentiations that all international surveys should seek to make (2003: 276; see also Table 3). Other writers have argued the theoretical pre-eminence of certain or multiple referents as markers of ethnic differences in the cross-national context (see Wrench & Solomos, 1993; Lloyd, 1995; Smith & Blanc, 1995; Rea et al., 1999; Aspinall, 2002).

---

3 Typically however, consensus on a choice of categories has been elusive. See for instance critiques of the British official classification (Ballard, 1997; Aspinall, 2002, 2003; Mason, 2003) – which arise despite extended consultations with social science practitioners during their development (Sillitoe & White, 1992; Owen, 1996).

4 The comparability problem is not unique to cross-national research. Green and Owen (1995), for instance, highlight the same issues with regard to regional structures in the UK.

5 As an example, contrast Britain, where measures of religion in combination with information on subjective ethnic identify are often argued to be the most significant markers of ethnic difference (Modood et al., 2002), with the US, where the importance of diversities between immigration cohorts and their originating countries is given most prominence (Alba & Nee, 2003)

Such attempts to prioritise specific ethnic referents can be presented as 'absolutist' approaches to the comparative measurement of ethnic difference. A common absolutist strategy in cross-national survey research is to focus upon only a single ethnic referent. Example choices involve immigrant or citizenship status (van Tubergen et al., 2004; SYSDEM, 2003); language minorities (Chiswick & Miller, 1995); or consideration of only certain distinctive ethnic groups (Model et al., 1999; Brown, 2000; Crul & Vermeulen, 2003).

Whilst such singular approaches offer an appealing conceptual clarity, they also do a disservice to the tremendous complexity of most countries' ethnic mosaics. An alternative absolutist approach involves choosing ethnicity categorisations which are defined by boundaries across several ethnic referents. The vast sociological literature on the interdependence and complexity of multiple alternative ethnic referents suggests that such a strategy is essential: no single referent can reasonably be studied out with the context of several others (e.g. Modood et al., 1997; Alba, 2005)[6]. Selected operationalisations of singular and cross-cutting absolutist ethnic categorisations in the cross-national context are discussed below.

The search for 'absolutist' comparability in categorisations of ethnic difference may, however, reflect only one approach to the cross-nationally comparative measurement of ethnic difference. An alternative form of comparability may be described as 'relativist'. Under this approach, particular ethnic categories within any country may be assigned a location within a dimension of difference which is defined relative to the structure of ethnic inequalities within the particular country[7]. Such relativist approaches to comparable measurement have been employed in other fields of cross-national survey research – for instance, measures of income, occupation or education standardised according to national averages (cf. Hoffmeyer-Zlotnik & Wolf, 2003). Relativist measures have not been widely considered in the context of measures of ethnicity, although recent papers by Lambert and Penn (2001), and Alba (2005), have taken this stance[8]. We argue below that relativist measures of ethnic difference have several attractions to cross-national survey researchers.

---

6 This same point may be phrased in terms of 'omitted variable bias' or 'spuriousness'. The exclusion of information on some relevant referents could lead to an impartial and misspecified account of the influence of others. Chiswick and Miller (1995), for instance, illustrate that information on language ability (referent 1) can be a highly relevant explanatory variable in studies of the economic impact of immigrant status (referent 2).

7 Hoffmeyer-Zlotnik & Wolf (2003) note how 'functional equivalence' is required in order to make cross-nationally comparative statements. This equivalence requires that measurement from diverse countries always refers to the same dimension of difference. Both 'absolutist' and 'relativist' categorisations can be presented as functionally equivalent in this regard.

8 An earlier relativist proposal, applied specifically to Britain, was made by Prandy (1979). van Tubergen's (2004) analysis may also be characterised as a relativist approach, as it incorporates measures of immigration background which refer to the relative characteristics of the sending and receiving communities.

3 Measurement of Ethnicity within Contemporary Surveys

The briefest examination of existing survey resources quickly reveals the difficulties of obtaining cross-nationally harmonised information on ethnic differences. Tables 1 and 2 illustrate a number of patterns in the data resources of four cross-nationally harmonised survey data collections. These are the European Social Survey (ESS, see Jowell, 2003 and *www.europeansocialsurvey.org/*; the ESS data is accessed via the Norwegian Social Science Data Services); the International Social Science Project datasets (ISSP, see Braun & Uher, 2003 and www.issp.org ; the ISSP data was accessed via the UK Data Archive at the University of Essex); the World Values Survey datasets (WVS, see Inglehart, 2000 and *www.worldvaluessurvey.org/*; the WVS data was accessed via the UK Data Archive at the University of Essex); and the studies of the Luxembourg Income Study (LIS, see an earlier review as Lambert & Penn, 2001, or *www.lisproject.org*; the LIS may be accessed direct via the project webpages). The survey collections differ considerably in their data collection strategies. The ESS attempts to follow the highest standards of 'pre-harmonisation' of questions before going to field, resulting in high levels of consistency between questions across different countries. Both the WVS and ISSP similarly use pre-harmonisation techniques, although more flexibility between countries and time periods is built into their designs. Lastly the LIS surveys are entirely 'ex-post' harmonised, one consequence being more examples of incomparable questions between countries, as well as higher risks of coding and translation errors (e.g. van Deth, 2003)[9].

Tables 1 and 2 can be summarised by two concise points[10]. The first (see Table 1) is that most of the surveys, and particularly the ESS, have nominally a very good collection of ethnicity related data, covering several alternative ethnic referents. The catch, however (see Table 2), is that the uses to which that source data could be used are highly constrained, given the sparsity of the responses involved.

9 Many other surveys may have been considered, though these four studies illustrate very typical properties of the medium-scale national sample surveys which are perhaps most widely used in cross-national survey comparisons.

10 Fuller elaborations of these tables, breaking down many of the components by particular countries, may be obtained on request from the author (also www.staff.stir.ac.uk/paul.lambert/downloads.html).

Table 2 'Wealth' of Ethnicity Data on 4 Survey Collections

(Ethnic referent codes as Table 1)

#Cat Number of categories in original data
#NSC Number of non-sparse categories (more than 50 cases, absolute value)
Skew Skewness: Percent of valid cases in the largest category
%m Percent of cases with missing data

ESS 2002: pooled data, 21 countries

	#Cat	#NSC	Skew	%m
	Country of citizenship			
C	118	5	96	0
	Country of birth			
B	158	12	90	0
	Time living in country (categories)			
T	6	5	90	0
	Language spoken at home			
L	102	9	95	1
	Whether in a minority ethnic group			
E	2	2	94	2
	Parents national origins (continent of parents' birth / mixed parentage)			
P	7	6	84	0
	Religious denom. (current or past)			
R	9	9	40	1

ESS 2002: Estimates are absolute values for merged dataset. Typical n = 2100 cases / country.

ESS Data – illustrative countries:

UK (GB), n=2052

	#Cat	#NSC	Skew	%m
C	28	1	97	0
B	57	1	91	0
T	6	2	91	0
L	31	1	96	0
E	2	1	94	0
P	7	3	85	1
R	9	4	45	0

Germany, n=2919

	#Cat	#NSC	Skew	%m
C	43	1	96	0
B	51	1	93	0
T	6	2	93	0
L	20	1	96	0
E	2	2	96	0
P	6	3	86	0
R	9	5	38	1

ISSP 2000: pooled data, 28 countries

	#Cat	#NSC	Skew	%m
C	10	1	95	0
B	20	2	80	2
L	4	2	60	0
E	10	2	80	20
R	8	3	50	2

ISSP: Estimates are *typical* values per country
Typical n = 1000 cases / country

WVS 1995/7: pooled data, 20 countries

	#Cat	#NSC	Skew	%m
B	100	12	91	2
T	7	7	92	21
L	100	12	95	9
E	200	10	90	14
R	14	11	38	4

WVS 1995-7: Estimates are absolute values for merged dataset. Typical n = 1000 cases/country

LIS c1995: pooled data, 24 countries

	#Cat	#NSC	Skew	%m
C	50	2	90	1
B	50	4	90	0
T	m	m	90	1
L	2	2	94	1
E	10	3	95	2
P	7	5	90	0

LIS: Estimates are typical values per country
Typical n = 50000 cases / country

However, closer inspection of the data from any particular study quickly reveals the superficiality of many of the apparently complete records. Table 2 summarises the 'wealth' of the ethnic referent data across the surveys[11]. Almost all of the differentiations recorded are categorical in nature, and the first column [#Cat] indicates the number of unique categories measured by the relevant variables. The second and third columns then indicate the sparsity and skew of the relevant distributions. Skewness, as the proportion of cases clustering into the largest category [Skew], is an indicator of how much variation is likely to be usefully analysed – a highly skewed variable can offer little differentiating information between cases, especially if the overall sample size is relatively small. The sparsity measure [#NSC] checks the absolute number of cases in relevant ethnicity categories, listing the number of categories with more than 50 people representing them. This again indicates how much analysis can realistically be undertaken on the variables – the absolute number of cases is ultimately more important than skew in this regard. Lastly, the fourth column [%m] shows the number of missing cases for each relevant variable – in some literatures it is generally expected that questions relating to ethnicity will be characterised by high levels of missing data, though this is not generally borne out across the range of datasets shown.

Table 2 summarises these patterns of data 'wealth' for the pooled cross-national samples (ie, with data from different countries combined). Although crude[12], this serves to illustrate the key features of the survey data. This is of a clear pattern of 'impoverished' data resources on ethnicity for the majority of countries. Although missing data is minimal, the data distributions are highly sparse and skewed, so that there are in practice likely to be very few circumstances where the divisions will sustain an informative analysis of ethnicity.

Compounding these basic distributional weaknesses, closer inspection of many of the datasets reveals that they often fail to measure some of the more theoretically appealing features of ethnic difference. For instance, few studies have extended data on subjective ethnic identity divisions (the ESS for example employs a dichotomous record of whether or not the subject identifies with any minority group). Few countries have substantial data on parental place birth (for the ESS, the data only differentiates 5 alternate continents). And the apparent 'wealth' of responses to questions on religion turns out in most cases to hinge on substantively less interesting divisions between categories of 'no religion' and one or more major Christian churches.

11 We use the term 'wealth' in reference to survey analysts' common descriptions of the 'richness of the data' : wealthy data contains high variability across cases on the relevant variables, and thus 'relationships between variables have the maximum chance to show up' (Punch, 2003: 38).

12 See Footnote 10

The relatively small samples of the ESS, ISSP and WVS offer further problems of data wealth. However it can be noted that within each of these studies, some specific countries do exhibit greater variation across ethnicity measures in a 'wealthier' way. These three studies are also attractive to researchers because the elements of pre-harmonisation in their design make data access and documentation relatively easy[13]. Although the LIS data studies have similar levels of skew in their relevant distributions, it can be noted that their number of cases in often much larger, thus meaning that sparsity of representation of categories can be less of a problem. The LIS data may also be noted for the wide range of ethnic referents measured between countries. This reflects greater national specific variations in data collection. This contributes both to conceptual confusion, but also to increasing the degree to which the LIS resources engage with ethnic referents of national specific research concern.

The picture of ethnicity information available on these four groups of surveys is messy and problematic. It should be remembered, moreover, that harmonised survey resources like these represent the stronger examples of comparative survey resources. Other researchers have reported greater difficulties attempting to conduct comparative analysis involving ethnicity on survey data collections which have not been subjected to the same levels of harmonisation and documentation (for instance, Stille, 1999).

4 Absolutist Solutions to Cross-National Research on Ethnicity

As Tables 1 and 2 have illustrated, the type of ethnic referent data collected by national surveys is inconsistent, and its analysis problematic due to the sparse numbers of cases representing many of the minority categories. An 'absolutist' perspective to the harmonisation of ethnicity data, therefore, suggests a pragmatic strategy, of choosing categories which may be both readily operationalised in most survey collections, and which will have modest numbers of cases in their minority categorisations. We list below six 'absolutist' categorical definitions of ethnicity differences which generate at least moderately 'wealthy' data structures for the survey datasets under study:

(1) [IMM] Immigrant status: a dichotomy indicating whether or not a case was born in the current country. This measure is widely used in previous research (though many studies record the age at entry to the host country). However it conveys limited information about ethnic differences, as it fails to make internal differentiations

13 A drawback in working with the WVS data on ethnicity is that the level of documentation in the centrally distributed files can be somewhat limited: readers must cross-check between multiple documents to obtain pithy descriptions of category labels, and, in a few example countries, apparent ethnic group data has no accessible documentation pertaining to it (thus excluded from Tables 1 and 2).

between different immigrant backgrounds, and it cannot recognise non-immigrant minorities, or even the children or grandchildren of immigrant ancestors (e.g. Banton, 1997). It is also flawed because of its popular conflation with concepts of citizenship – Hoffmeyer-Zlotnik (2003) emphasises how the two measures are analytically distinct).

(2) [LAN] Minority language use: a dichotomy indicating whether or not a case generally speaks a language other than an official majority language of the host country. This measure has considerable sociological significance, as the analysis of ethnic differences is increasingly concerned with information on language use (e.g. Portes & Rumbaut, 2001; Alba & Nee, 2003). However this measure has several flaws: there is a lack of cross-national consistency to the measurement of 'minority' languages, and a subjectivity to individuals' reports of usage. It also ignores ethnic minorities who do not use minority languages.

(3) [VIS] Visible minority group status: a dichotomy indicating whether or not a case belongs to a minority group on the grounds of any overt ethnic group formation – such as racialised visibility, subjective ethnic group identity, or participation in a 'visible' minority religion[14]. This measure highlights the ethnicity referents which are most popular with leading sociological thinking in the field (cf. Modood, 2002); however, it conceals considerable internal heterogeneity.

(4) [MIN] Any minority group membership: a dichotomy which extends the VIS categorisation to highlight any ethnic minority identity from any relevant referent – citizenship, country of birth, parental national origins, language use, and visible minority group status. Again, this catholic measure masks diversity within its categories. It is also unlikely that the same differentiations will contribute to the same categories in different countries.

(5) [CON] National-specific scheme: a multiple categorical scheme chosen from the available schemas favoured by national literatures. This has clear substantive attractions, but is problematic for comparative purposes.

(6) [EC9] Comparative ethnicity 9 category measure: a multiple categorical scheme advocated here as an attempt at incorporating as many ethnicity differentiations (cf. Hoffmeyer-Zlotnik, 2003) as reasonably possible. It uses information from five dichotomous measures, IMM, LAN, VIS, and indicators of citizenship and of whether or not either parent was born outside the country. That information is cross-classified to form the categories listed below, where 'minority group' corresponds to either VIS=1, or having non-host citizenship:

14 This religion criteria used here reflects the perspective (and approximation), advocated for instance by Modood et al. (2002) and Brown (2000), that only certain minority religions are most important to social stratification outcomes in the contemporary societies under study: here defined as recording a South Asian or Islamic religious identity.

1 CCNN – Self and parents born in country, no minority group or language
2 CCMN – Self and parents born in country, minority group but not language
3 CCL– Self and parents born in country, minority language
4 CPNN – Self born in country, parents not, no minority group or language
5 CPMN – Self born in country, parents not, minority group but not language
6 CPL – Self born in country, parents abroad, minority language
7 FNN – Self born abroad, no minority group or language
8 FMN – Self born abroad, minority group but not language
9 FL – Self born abroad, minority language

Table 3 summarises the distribution of cases to these categorical schemes from some of the survey collections considered. Its distributions are indicative of a number of important points. Firstly, the four dichotomous measures – whilst distinctly unsatisfactory as a theoretical categorisation – have more attractive data distributions. On the other hand, the national specific categorisations remain highly problematic – the skewness and sparsity of their minority categorisations is pronounced, whilst there is inconsistency in the number of categories delineated between countries. Lastly, the proposed 'EC9' scheme appears to be reasonably appropriate. It largely avoids sparsity problems (with the exception of the CCL and CPL categories, minority groups with regard to language but not other criteria), and is in most circumstances capable of being derived for most countries of interest (with some exceptions).

The value of these alternative categorisations can also be assessed by empirical judgements on the strength of association between ethnicity indicators and other social measures. When we analyse, for example, indicators of social stratification advantage, we see clear patterns of association with these six ethnicity measures. Importantly, the four dichotomies (particularly the 'MIN' categorisation), prove relatively strong, accounting for the large bulk of outcome variation that can be attributed to further ethnicity differentiations. Also of interest is the observation that slightly more variation is typically associated with the EC9 measure (which has no national specific input), than with the CON measure. However, other analyses of these ethnicity indicators leads to less favourable conclusions. We frequently observe non-significant main, and interaction, effects associated with dummy variable indicators of ethnic minority groups (variously defined), when used as predictors of social stratification advantage in human capital style regression models. On theoretical grounds we have strong reasons for expecting such associations to exist – it can be assumed that the small number of cases in many minority groups is making evidence of their effects difficult to confirm.

Table 3 'Absolutist' Ethnicity Measures: Properties of Alternative Categorical Differentiations (n's representing ethnic categories, ESS/WVS/ISSP data)

	ESS	WVS
IMM *(if born in host country)*		
0	36839	21736
1	3935	2129
LAN *(if speak minority language)*		
0	36383	20311
1	4336	3516
VIS *(if a 'visible' minority group)*		
0	36261	16655
1	3655	5236
MIN *(if minority by any referent)*		
0	29176	14015
1	10716	6886

cf.: Hoffmeyer-Zlotnik (2003: 276)
1 – Legal status (citizenship)
2 – Country of birth
3 – Visibility
4 – Mother tongue
5 – Cultural customs / religion
6 – Self-perception

	ESS	WVS	ISSP UK	ISSP Germ.
CON *(national specific ethnicity categories)*				
1			919	1417
2	n/a	n/a	12	20
3			7	10
4			9	15
5			5	37
6			2	
7			1	
8			3	
9			13	
EC9 *(minority categories by cross-classification of referents)*				
1 – CCNN	29176	1405	1642	2432
2 – CCMN	832	2048	60	47
3 – CCL	547	1783	1	1
4 – CPNN	2051		92	161
5 – CPNM	850		34	19
6 – CPL	122		5	11
7 – FNN	1205	597	63	59
8 – FMN	1216	292	68	64
9 – FL	1115	847	50	69
% missing	9.2	25.5	1.8	1.9

This evaluation of 'absolutist' representations of ethnic differences helps illustrates several issues. Firstly, a number of dichotomous measures of concepts are readily operationalised in cross-national surveys and exhibit discernible associations with other relevant factors; they may also be easily understood and communicated. They are, however, compromised by substantial internal heterogeneity, which may mask otherwise significant ethnic differences. Alternatively, national specific measures ('CON' variables) are seriously constrained, by a lack of conceptual comparability between countries; by the lack of availability of the relevant data for many surveys; and by the sparse representation of many contributing categories. Lastly, the 'EC9' categorical representation has both conceptual attractions (as a consistent measure between countries) and weaknesses (its possible disengagement from national-specific concerns over ethnicity differences). As a practical measure EC9 has intermediate properties. Its association with other variables is particularly strong, and in most circumstances there is a moderate spread of cases between different categories. However, categories of the measure still tend towards sparsity in certain contexts, whilst many of its categories continue to conflate sociologically signifi-

cant ethnicity differences. It may thus be argued that with harmonised social surveys, it remains difficult to devise and analyse satisfactory categorical ethnicity classifications which apply to national populations in a meaningful way.

5 Relativist Solutions to Cross-National Research on Ethnicity

A 'relativist' approach to recording ethnic differences involves two stages. The first is to record ethnic locations, as mapped by information on multiple possible ethnic referents. The second is to analyse that location, not simply as a qualitative category, but as a position within a dimension of relative ethnic difference. This is often achieved mathematically, by assigning metric scores to different categorical locations (for earlier developments of this argument, see Lambert & Penn, 2001).

Table 4 gives an illustrative selection of such derived ethnic category scores. The columns of Table 4 show the results from a series of Stereotyped Ordered Regression (SOR) models (Anderson, 1984). Here, ethnic category scores are derived as a function of average differences between ethnic categories according to a variety of predictor variables[15]. Metric scores derived by SOR models for a selection of EC9 and CON categories are shown: the values represent a dimension of difference between categories, as they are associated with a regression format model summarising demographic and economic effects. The precise interpretation of the scores hinges upon the parameter coefficients of the regression models on which they are derived (not shown), but the important point is that the SOR scores represent average relative differences between individuals from different ethnic categories in a country along a dimension of measured social and economic circumstances.

In this case, a scaling is thus derived which reflects *relative* ethnic difference within each country. In practice, such SOR scores could be derived from national specific information sources (with greater coverage of cases), then mapped back to the sparser categories identified in sample surveys. In principle, SOR scores could be estimated for the dozens of different categories which emerge when cross-classifying multiple ethnic referents. The subsequent metric treatment of these categories can then be exploited to avoid the problems of undertaking categorical data analyses techniques with sparsely represented groups.

15 The SOR models were derived using macros following Hendrickx (2000); for discussion of the specification of these models for this context, see Lambert and Penn (2001).

Table 4 'Relativist' Ethnicity Measures: 'SOR' Score Derivations for Selected ESS Samples

SOR regression: *[Ethnicity] = gender + age + education + marital status + employment status + employment advantage + interactions*

	North-West Europe sample, n=20899		UK only, n=1893		Germany, n=2503	
SOR scores for EC9	**EC9**	**EC9**	**EC9**	**CON_U**	**EC9**	**CON_G**
1. CCNN	-239	-530	-253		-332	
2. CCMN	-220	-144	-152		-369	
3. CCL	-380†	-368†	-	-571	293†	-504
4. CPNN	-118	-22	-226	309†	-328	826†
5. CPMN	233	192	-143	643†	461†	-235
6. CPL	634	604	717†	22	324†	-87
7. FNN	-330	-203	484	-404	-387	
8. FMN	5	171	-241		33	
9. FL	414	300	-186		305	

† : Less than 20 cases in the SOR category
CON_U: 1=White; 2=Black-Caribbean; 3=Black-African; 4=Asian; 5=Other.
CON_G: 1=German citizen; 2=Turkish citizen; 3=Eastern European citizen; 4=Other citizen.

There are of course many other ways in which ethnic categories could reasonably be assigned score values. For instance this could be through using other summarising functions which have perhaps simpler and more communicable interpretations (say the average employment advantage score associated with each group), or perhaps by *fiat*, allowing sociological thinkers an opportunity to make judgements reflecting a wider range of considerations concerning ethnic differences. However derived, category scoring offers an opportunity to analyse ethnicity differences in hierarchical manners, interpreted as the typical effects of ethnicity differences as they operate through a given structure of inequalities[16].

There are several attractions to this strategy. The metric representation can bypass many data analytical difficulties associated with the low representation of minority group categories. The metric representation can itself prove a highly revealing communication on the nature of ethnic differences within any given country. A relativist metric can chime more closely with recent sociological theorising on the complexity, diversity, dynamism and contextuality of ethnic differences, as its flexibility can allow the mapping of multiple ethnic locations defined in terms of numerous cross-cutting referents. Lastly, it can be

16 It is important to note that metric variable representations should not be conflated with unidimensionality – multiple dimensions of differences can be incorporated (Lambert & Penn, 2001).

argued that relative measures of ethnic differences have a form of cross-national comparability so long as the underlying structure of ethnic differences is estimated in a similar way across countries – in the same fashion, for example, as a measure of income that is standardised around national averages.

The potential value of such ethnic category scoring is that it may parsimoniously summarise the key elements of ethnic differences that affect the population under study. Empirical analyses have so far only partially supported this view. For instance, the derived scores presented in Table 4 exhibit only modest patterns of association with other variables which are expected to relate to ethnic differences. Lambert and Penn (2001) similarly found only a few circumstances where SOR score effects altered the estimates from categorically based analytical regression models. Nevertheless, there is a strong possibility that a more powerful relativist scoring framework could be uncovered, and thus that the assignment of ethnic category scores could prove a significant benefit to cross-national analysis.

6 Conclusions

This review has identified three significant problems in the analysis of ethnicity in cross-national studies. Firstly, it has shown that existing surveys resources are haphazard in their measurement of ethnic referents between different countries and time periods. Certain more consistent choices in data collection strategy could be employed (e.g. Hoffmeyer-Zlotnik, 2003), though to date, the European Social Survey is the only major harmonised collection which has come close to achieving these standards. Secondly, we have seen that regardless of the range of relevant questions covered in surveys, in almost all instances, harmonised survey collections suffer severe problems of data 'wealth' with regard to ethnicity information. This is primarily a function of the sparse representation of the minority groups. Thirdly, this review has also demonstrated severe theoretical critiques of the cross-national analysis of ethnicity effects. These focus upon the apparent inability of harmonised categorisations to reflect the full and appropriate 'ethnic mosaic' of the societies under study. However, it is argued that despite such problems, we should not shy away from the analysis of ethnicity in a comparative perspective, and this review has sought to outline reasonable solutions from two alternative perspectives.

A 'relativist' approach to the harmonisation of ethnicity data involves categorisations, from whichever ethnic referent is available, being mapped onto one or more dimensions of ethnic inequality, and given numerical indicator scores for their relative positions. This strategy is cognitively challenging, as it requires the researcher to communicate abstract ideas on relative positions within a hypothesised social structure. It is also empirically

ambiguous, as it is only in selected circumstances that relativist measures of ethnic difference appear to exhibit greater criterion validity that absolutist measures. However, relativist measures have considerable theoretical appeal, since they are far more amenable to cross-national (and temporal) variations in the type of ethnic data collected, and offer an abstract form of functional equivalence which exceeds those of absolutist schemes.

'Absolutist' approaches to harmonisation, which involve ensuring that the same concepts are used to differentiate categories in different countries, are more easily communicated, and may be expected to continue to be widely used. This strategy can lead to crude measures of ethnic difference, perhaps involving only one ethnic referent or the merger of a number of loosely related minority group categories. However it can be argued that certain carefully applied absolutist categorisations can be more satisfactory. The 'EC9' categorisation illustrated above, for instance, has reasonably strong analytical properties and is apparently appropriate for the majority of countries considered in this review.

References

Ahmad, W. 1999. "Ethnic Statistics: Better than nothing or worse than nothing?" Pp. 124-131 in *Statistics in Society: The Arithmetic of Politics* edited by Dorling, D., and S. Simpson. London: Arnold.

Alba, R. 2005. "Bright vs. blurred boundaries: Second-generation assimilation and exclusion in France, Germany, and the United States." *Ethnic and Racial Studies* 28:20-49.

Alba, R., and V. Nee. 2003. *Remaking the American Mainstream: Assimilation and Contemporary Immigration.* Harvard: Harvard University Press.

Allen, S., and M. Macey. 1990. "Race and Ethnicity in the European Context." *British Journal of Sociology* 41:375-393.

Anderson, J. A. 1984. "Regression and Ordered Categorical Variables." *Journal of the Royal Statistical Society Series B* 46:1-30.

Anthias, F. 2001. "The concept of 'social division' and theorizing social stratification: Looking at ethnicity and class." *Sociology* 35:835-854.

Aspinall, P. J. 2002. "Collective terminology to describe the minority ethnic population: The persistence of confusion and ambiguity in usage." *Sociology* 36:803-816.

Aspinall, P. J. 2003. "The conceptualisation and categorisation of mixed race/ethnicity in Britain and North America: Identity options and the role of the state." *International Journal of Intercultural Relations* 27:269-296.

Ballard, R. 1997. "The construction of a conceptual vision : Ethnic groups and the 1991 UK census." *Ethnic and Racial Studies* 20:182-194.

Banton, M. 1997. *Ethnic and Racial Consciousness.* London: Longman.

Bonifazi, C., and S. Strozza, (Eds.). 2003. *Integration of migrants in Europe: data sources and measurement in old and new receiving countries.* Special issue of International Journal of Migration Studies 40(152).

Braun, M., and R. Uher. 2003. "The ISSP and its Approach to Background Variables." Pp. 33-48 in *Advances in Cross-National Comparison,* edited by Hoffmeyer-Zlotnik, J. H. P., and C. Wolf. New York: Kluwer.

Brown, M. S. 2000. "Quantifying the Muslim population in Europe: conceptual and data issues." *International Journal of Social Research Methodology* 3:87-102.

Castles, S. 1993. "Migrants and Minorities in Europe. Perspectives for the 1990s: Eleven Hypotheses". Pp. 17-34 in *Racism and Migration in Western Europe,* edited by Wrench, J., and J. Solomos. Oxford: Berg.

Chiswick, B. R., and P. W. Miller. 1995. "The Endogeneity between Language and Earnings – International Analyses." *Journal of Labor Economics* 13:246-288.

Crul, M., and H. Vermeulen. 2003. "The Second Generation in Europe." *International Migration Review* 37:965-986.

Evans, G., and A. Need. 2002. "Explaining ethnic polarization over attitudes towards minority rights in Eastern Europe." *Social Science Research* 31:653-680.

Favell, A. 2001. *Philosophies of Integration: Immigration and the Idea of Citizenship in France.* 2nd Edition. Palgrave: Macmillan.

Favell, A. 2003. "Integration Nations: The Nation-State and Research on Immigrants in Western Europe." *Comparative Social Research* 22:13-42.

Green, A. E., and D. Owen. 1995. "Ethnic minority groups in regional and local labour markets in Britain: A review of data sources and associated issues." *Regional Studies* 29:729-735.

Harkness, J. A., F. J. R. van de Vijver, and P. P. Mohler. 2003. *Cross-Cultural Survey Methods.* New York: Wiley.

Heckmann, F., R. D. Penn, and D. Schnapper. 2001. *Effectiveness of National Integration Strategies Towards Second Generation Migrant Youth in a Comparative Perspective – EFFNATIS.* Bamberg: European Forum for Migration Studies, University of Bamberg.

Hendrickx, J. 2000. "Special Restrictions in multinomial logistic regression." *Stata Technical Bulletin* 56:18-26.

Hoffmeyer-Zlotnik, J. H. P. 2003. "How to Measure Race and Ethnicity" Pp. 267-277 in *Advances in Cross-National Comparison: A European Working Book for Demographic and Socio-Economic Variables,* edited by Hoffmeyer-Zlotnik, J. H. P., and C. Wolf. New York: Kluwer Academic.

Hoffmeyer-Zlotnik, J. H. P., and C. Wolf (Eds.). 2003. *Advances in Cross-National Comparison: A European Working Book for Demographic and Socio-Economic Variables.* New York: Kluwer Academic.

Hutchinson, J., and A. D. Smith (Eds.). 1996. *Ethnicity.* Oxford: Oxford University Press.

Inglehart, R. 2003. *World Values Surveys and European Values Surveys* 1981-4, 1990-3, 1995-7 [Computer file]. Ann Arbor, MI : Institute for Social Research [Producer]; Ann Arbor, MI : Inter-university Consortium for Political and Social Research [Distributor]. 26.7.02.

Jacobs, D., and J. Tillie. (Eds.). 2004. "Social Capital and Political Integration of Migrants." *Journal of Ethnic and Migration Studies* 30:419-427.

Jowell, R. 2003. *European Social Survey 2002/2003: Technical Report.* London: Centre for Comparative Social Surveys, City University.

Lambert, P. S., and R. D. Penn. 2001. *SOR models and Ethnicity data in LIS and LES : Country by Country Report.* Syracuse University, Syracuse, New York 13244-1020: Luxembourg Employment Study Paper No. 21, Maxwell School of Citizenship and Public Affairs. (http://www.lisproject.org/wpapers.htm).

Lloyd, C. 1995. "International Comparisons in the Field of Ethnic Relations." Pp. 31-46 in *Racism, Ethnicity and Politics in Contemporary Europe,* edited by Hargreaves, A. G., and J. Leaman. Aldershot: Edward Elgar.

Model, S., G. Fisher, and R. Silberman. 1999. "Black Caribbeans in comparative perspective." *Journal of Ethnic and Migration Studies* 25:187-212.

Modood, T., R. Berthoud, J. Lakey, J. Y. Nazroo, P. Smith, S. Virdee, and S. Beishon. 1997. *Ethnic Minorities in Britain: Diversity and Disadvantage.* London: Policy Studies Institute.

Modood, T., R. Berthoud, and J. Nazroo. 2002. " 'Race', racism and ethnicity: A response to Ken Smith". *Sociology* 36:419-427.

Owen, D. 1996. *Towards 2001: Ethnic Minorities and the Census.* Warwick: Centre for Research in Ethnic Relations, University of Warwick.

Portes, A., and R. G. Rumbaut. 2001. *Legacies: The Story of the Immigrant Second Generation.* Berkeley and Los Angeles: University of California Press.

Prandy, K. 1979. "Ethnic discrimination in employment and housing." *Ethnic and Racial Studies* 2:66-79.

Punch, K. F. 2003. *Survey Research: The Basics.* London: Sage.

Rea, A., J. Wrench, and N. Ouali. 1999. "Discrimination and Diversity." In *Migrants, ethnic minorities and the labour market : Integration and Exclusion in Europe,* edited by Wrench, J., A. Rea, and N. Ouali. London: Palgrave.

Sanders, J. M. 2002. "Ethnic boundaries and identity in plural societies." *Annual Review of Sociology* 28:327-357.

Sillitoe, K., and P. H. White. 1992. "Ethnic Group and the British Census : the search for a question." *Journal of the Royal Statistical Society, Series A: Statistics in Society* 155:141-163.

Smith, D. M., and M. Blanc. 1995. "Some Comparative Aspects of Ethnicity and Citizenship in the European Union." Pp. 70-92 in *Migration, Citizenship and Ethno-National Identities in the European Union,* edited by Martiniello, M. Aldershot: Avebury.

Stille, F. 1999. *Ethnic Minorities and Immigrant Groups on the Labour Market : Introduction.* SYSDEM: Employment Observatory Trends 32.

SYSDEM. 2003. "Dealing with the Challenge of Immigration." *European Employment Observatory Review.* Autumn, 27-112.

van Deth, J. W. 2003. "Using Published Survey Data." Pp. 329-346 in *Cross-Cultural Survey Methods,* edited by Harkness, J. A., F. J. R. van de Vijver, and P. Ph. Mohler. New York: Wiley.

van Tubergen, F., I. Maas, and H. Flap. 2004. "The Economic Incorporation of Immigrants in 18 Western Societies: Origin, Destination and Community Effects." *American Sociological Review* 69:704-727.

Vertovec, S. 2003. "Migration and other modes of transnationalism: Towards conceptual cross-fertilization." *International Migration Review* 37:641-665.

Wrench, J., and J. Solomos. 1993. *Racism and Migration in Western Europe.* Oxford: Berg.

MEASURING RELIGIOUS AFFILIATION AND RELIGIOSITY IN EUROPE

CHRISTOF WOLF

1 Introduction

When Constantine declared Christianity to one of the official religions of the Roman Empire in 313 AD Europe acknowledged her beginning Christian tradition. At that time approximately 15% of the Empire's population were Christians. Only 67 years later in 380 AD Emperor Theodosius I declared Christianity to the sole state religion. Not surprisingly what we think of as Europe now was referred to as "Christendom" at least between the 14th and 18th century (Davies, 1996: 7).

Within Europe's Christendom vast differences between different flavors of Christianity often leading to deadly conflict and war were and to a certain degree still are present (Pelikan, 1987; Harakas, 1987). In addition, Europe always was populated by large minorities adhering to other religions; mostly Jews and Muslims. More recently we experienced the spread of Asian religions, too. Europe is also marked by a high degree of variation of the relationship between state and church (Robbers, 1996). In some countries State Churches can be found while others are marked by the strict separation of state and church. Overall, the religious landscape in Europe is marked by diversity, partly reflected by differences between countries and regions.

Given this diversity the question I address in this paper is whether we are able to measure religiosity across Europe. Where the instrument should be applicable in large-scale multipurpose surveys and thus has to be economical. Two approaches to this kind of measurement are investigated. First, membership in or identification with religious groups are studied. Second, religious behaviors and attitudes are considered. Both types of approaches are empirically analyzed on the basis of two international comparative surveys: The International Social Survey Programme (ISSP) on the topic of religion from 1998 and the first round of the European Survey (ESS) from 2002. To make these datasets as comparable as possible only data from the 18 countries which where included in both studies are

used in the following analysis (cf. Table A1 in the appendix).[1] For the analysis presented in this paper the design weights supplied with the datasets were always applied. Where summary results for 'Europe' are reported an additional population weight was used. All references to Europe or Europe's population refer only to the countries selected for the analysis.

It should be noted that the ISSP and the ESS differ in several respects. First, partly as a reaction to the experience made in the context of the ISSP which is an ongoing project since 1985, the ESS is much more centrally organized and presents strict guidelines to all participants regarding survey procedures. The ISSP applies less rigid procedures, party to save costs and to be able to adapt to national survey customs. For this reason the ISSP has deliberately renounced using a common source questionnaire for the background variables such as education or income, but also religious affiliation and church attendance. Thus, the way these data are collected varies between countries and a strategy of ex-post harmonization is applied to them(see Scholz, in this volume). In contrast the ESS master questionnaire encompasses all questions and only leaves the collection of educational standing open to national practice (Kolsrud & Skjåk, in this volume). A further important difference pertains to the mode of data collection. While the ISSP is usually administered as a written questionnaire often distributed in the context of some other national survey the ESS data are collected by face-to-face interviews. A minor difference are the differing age limits of the target population. This was set at 15 years of age for the ESS, though some countries have also sampled 14 year old respondents. In the ISSP the lower age limit seems to be 18 years, however, there are countries in which 16 years was set at the minimum age.

2 Measuring the Association with a Religious Group: Membership vs. Identification

Measuring the relationship between an individual and a religious group can be done in at least two different ways. First, one can ask if a person is a *member* of a religious group. Second, one can measure if a person *identifies* him- or herself with a religious group. Of course, to yield different results with these approaches the concept of *membership* has to be unambiguous, not left to individual interpretation. This condition may hold more true in some countries or with respect to some religious groups than in others. For example, in Germany membership in the mainline protestant churches as well as in the catholic church is indicated on the income tax card of employees, thus reminding people of their status.

1 Included in this number are East- and West-Germany (DO and DW respectively) which were treated as separate countries in the following analysis. The ISSP-data for North Ireland and Great Britain were combined to data for the United Kingdom to match the respective ESS data.

If we only distinguish between two states of membership and identification we can draw a four-fold table (see Table 1). There are two states in which both indicators are congruent: Either both signal an association with a religious group (religious member) or both signal the absence of such an association (secular non-member). Then there are two states in which the two indicators differ: Either members do not identify (nominal members) or non-members identify (religious non-members). Of course, these latter incongruent cases are especially interesting and the source of much debate among scholars of the sociology of religion. Proponents of the secularization thesis would argue that there is a general shift from the lower left corner of the table to the upper right corner with a detour via the nominal member state (Bruce, 2001; Demerath, 2001; for a definition see Wilson, 1987: 160). Opponents of this view claim that "many Europeans have ceased to connect with their religious institutions in any active sense, but they have not abandoned, so far, … their deep-seated religious aspirations …" (Davie, 2002: 8). According to this observation one would expect individuals to move from the left quadrant to the right quadrant of the bottom row; i.e. from being religious members to being religious non-members. The result of this process is characterized by Davie as "believing without belonging", an empirical description that was recently contested by Voas & Crockett (2005).

Table 1 A Typology of Membership and Identification

		Membership in religious group	
		yes	**no**
Identification with religious group	no	nominal member	secular non-member
	yes	religious member	religious non-member

Fortunately, there are data available to test the empirical validity of the two positions sketched above. At the same time these data allow us to explore the advantages and disadvantages of measuring association with religious groups based on the criteria of membership or identification. In the European Social Survey (ESS) association with religious groups is measured by the following question: "Do you consider yourself as belonging to any particular religion or denomination?" The stimulus "belonging" was explicitly chosen as an indicator for identification not membership (see ESS, round 1, source questionnaire, page 17).[2]

2 At least from a German perspective but certainly also from the perspective of other countries involved in the ESS this approach to measure religious affiliation is rather atypical. In Germany the question employed usually asks for "*membership* in a religious group". Consequently, in the German and Austrian versions of the ESS an additional clarifying sentence was added to the question given in the source questionnaire: "*Regardless of whether you are a member or affiliate of a church or religious group* do you consider yourself as belonging to any particular religion or denomination?".

Those giving a positive answer to this question were asked which group they feel they belong to. The source questionnaire lists 21 different groups five of which contain the addition "other", e.g. "Other Eastern Orthodox", that should be entered in detail (cf. Table 2).[3] As far as I can see this list was applied only in Great Britain. Most countries used far shorter lists, though at least one country, Ireland, asked for religious identification in an open ended question.

Table 2 List of Religious Groups Used in the ESS

Main Questionnaire		Round 1 Dataset	
01	Christian – no denomination		
02	Roman Catholic	1	Roman Catholic
03	Greek or Russian Orthodox	3	Eastern Orthodox
04	Other Eastern Orthodox, which		
05	Protestant (no further detail)	2	Protestant
06	Church of England / Anglican		
07	Baptist		
08	Methodist		
09	Presbyterian/Church of Scotland		
10	United Reform Church/Congregational		
11	Free Presbyterian		
12	Brethren		
13	Other Protestant, which		
14	Other Christian, which	4	Other Christian
15	Hindu	7	Eastern Religions
16	Sikh		
17	Buddhist		
18	Other Eastern Religions, which		
19	Jewish	5	Jewish
20	Islam / Muslim	6	Islam
21	Other non-Christian, which	8	Other non-Christian

In the integrated dataset only eight groups are distinguished (right column in Table 2). However, not even these groups can be identified in every country. The table of 9 religious groups – 8 groups plus those not belonging to any group – by 23 countries has 27 empty cells (13%). If we restrict the analysis to those cells with more than 10 cases 101 cells (49%) have to be discarded. The only group containing more than 10 cases in

3 An overview of different approaches to classifying religions can be found in Partin (1987).

every country included in the ESS is the group of those not identifying with any religious group. Furthermore, if a minimum of 10 cases per country is taken as the criterion, the category "Jews" can only be analyzed in Israel.

In contrast to the ESS the ISSP collects data on religious membership. Due to the fact that the data for the ISSP usually is collected in connection with other (national) surveys no fixed master questionnaire for the background variables exist. Instead the participating countries agree to collect data in such a way that allows them to code these to a predefined list of compulsory background variables and categories (cf. Braun & Uher, 2003).[4] Thus, an approach of output harmonization is applied. Because a mandatory question wording is missing in the ISSP and researchers instead rely on their country-specific conventions for gathering this information on association of the respondent with a religious group the exact meaning of this question might vary between countries. However, at least in the case of the European countries involved in the ISSP the predominant formulation is consistent with the interpretation that *membership* rather than *identification* is measured.

Because the ISSP is a world-wide survey carried out in almost 40 countries today the categories of religious groups and denominations proposed in the ISSP list of core background variables is a carefully designed hierarchical system resulting in a three-digit code that theoretically allows the distinction of hundreds of different groups. For ISSP data in which this classification is employed around 70 different categories are found.[5] However, the data that will be analyzed in what follows comes from the ISSP survey on religion from 1998. At that time only a little more than 20 religious groups and denominations were distinguished.

Table 3 presents the distribution for the two different concepts of measuring association with religious groups from the ISSP and the ESS. Overall, the two distributions show a high degree of similarity. With over 95% of the population associated with a religious group belonging to Christianity both datasets clearly reflect Europe's Christian tradition. Furthermore, both surveys show that the 'not affiliated' are the second largest 'religious' group in Europe. However, there are also striking differences between the two distributions: First, when asked if they *identify* with a religious group more than a third of

4 In the mandatory list of core background variables this variable is called RELIG and is described as "Religious denomination (asked country specific but re-coded to standard)" (Braun & Uher, 2003: 39). The ISSP list of background variables can also be attained at: http://www.za.uni-koeln.de/data/en/issp/codebooks/bv2001.pdf (2005/12/02).

5 The list of core background variables was agreed upon in 2001 and was applied for the first time in the data collection round of 2002.

Europe's population answers in the negative, compared with 'only' a quarter negating *membership*. Second, whereas the percentages of Catholics are virtually identical in the two datasets the proportion of the population claiming membership in a Protestant group is much higher than the proportion identifying with such a group. Third, in contrast to Protestants the proportion of the population identifying with small religious groups tend to be *larger* than the respective proportions identifying with these groups. This result could be an artefact resulting from differences in the way the data are collected in the two surveys and their slightly different definitions of the target populations. Nonetheless, this result could also reflect the small but noticeable increase in 'alternative' religions lacking traditional notions of membership, sometimes even lacking the concept of exclusiveness.

Table 3 Religious Membership/Association in Europe

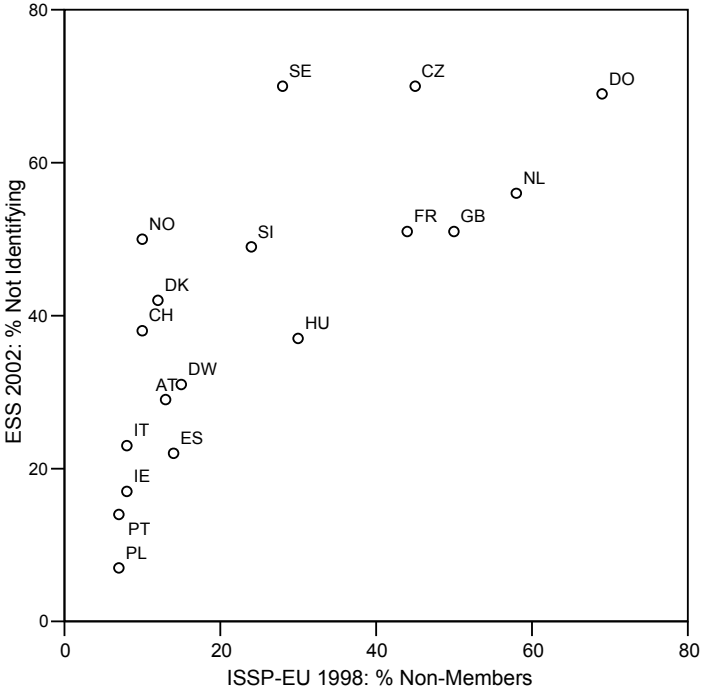
	ISSP-EU 1998			ESS 2002		
	A	B	C	A	B	C
Catholic	51.8	70.8	97.8	44.7	70.6	95.9
Protestant	19.1	26.2		14.4	22.7	
Eastern Orthodox	0.1	0.1		0.3	0.5	
Other Christian denomination	0.6	0.8		1.3	2.1	
Judaism	0.2	0.2	2.2	0.2	0.3	4.1
Islam	0.3	0.4		1.7	2.7	
Eastern religions	0.2	0.3		0.5	0.7	
Other non-Christian religions	0.9	1.2		0.2	0.3	
None	26.9	—	—	36.8	—	—
Unweighted N	22,008	16,098	16,098	31,610	19,052	19,052

A Percentage of total.
B Percentage of affiliated.
C Christian vs. non-Christian groups.

As can be seen from Figure 1 the differences between membership rates and the proportion identifying with a religious group differ greatly throughout Europe. In countries like the Netherlands, the United Kingdom, Hungary, Spain, Portugal, France and East-Germany the differences are modest. In Scandinavia the divergence between the two concepts is largest: In Denmark the discrepancy is 30 percentage points, in Sweden and Norway the difference amounts to around 40 percentage points.

There is however no clear relationship between the religious or denominational composition of a country and the degree to which membership and identification rates differ. There are protestant countries with large differences, e.g. Scandinavia, and protestant countries with no discrepancies, e.g. East-Germany. Nonetheless, the share of Catholics seems to be negatively related to divergence between the two measures under study. Thus, at least on the aggregate level belonging and identifying are more congruent in Catholic than in mixed or Protestant countries.

Figure 1 Identification vs. Membership



3 Indicators of Religiosity

Religion or rather religiosity is a multidimensional concept. At least since Glock's (1962) pioneering investigation into the measurement of religion we distinguish between religious belief, religious experience, religious knowledge and religious behavior. Where the latter might be divided into public religious behavior, e.g. church attendance, and private behavior, e.g. prayer.[6] Both, the ISSP 1998 and the ESS of 2002 contain several indicators related to these dimensions of religiosity. Three of these items were collected in very similar fashion in both surveys and are therefore suitable for comparison.

6 For the measurement of religiosity see Boos-Nünning (1972); Kecskes & Wolf (1993, 1995); Meulemann (1985); Schreuder (1991); Roof (1979); Steensland et al. (2000); Hill & Hood (1999); Slater, Hall & Edwards (2001).

Next to religious affiliation indicators measuring the frequency of worshiping are of central importance to the study of religiosity (Jagodzinski & Dobbelaere, 1995; Feldkirchner, 1998). Two of these indicators are included in the ISSP and the ESS: attending religious services and praying. Box 1 contains the questions and answering categories for collecting the data on the frequency of attending religious services employed in the ISSP and the ESS. In both instances the answering categories run from the more frequent to never, though there is some variation. The ESS categories differentiate more in the region of higher frequencies, whereas the ISSP measure is somewhat more fine grained in the middle region. Overall however the similarities prevail.

Box 1 Measuring Attendance at Religious Services

	ISSP 1998 [a]	**ESS 2002**
Question:	According to national practice, for example in France: En dehors des mariages, des enterrements et des baptêmes, tous les combien assistez-vous à un culte religieux ?	Apart from special occasions such as weddings and funerals, about how often do you attend religious services nowadays?
Answers:	1 Once a week or more 2 2-3 times a month 3 Once a month 4 Several times a year 5 Less frequently a year 6 Never	1 Every day 2 More than once a week 3 Once a week 4 At least once a month 5 Only on special holy days 6 Less often 7 Never

a The scheme for measuring attendance at religious services in the ISSP was updated 2001 by adding two categories: 'several times a week' and 'once a year'.

With respect to the measures of praying given in Box 2 the differences are more pronounced. The answering scales are reversed and they vary with respect to their degree of differentiation. In the ISSP an 11-point scale is used with categories running from 'never' to 'several times a day' and the corresponding numbers increase from '1' to '11'. The ESS uses the same scale as for the measurement of attending religious services, that is a 7-point scale reaching from 'every day' represented by a '1' to 'never' marked by the number '7'. Given findings from experimental studies on numerical coding and poling of answering scales it could be expected that the two questions vary in their relation to other indicators of religiosity (Krebs & Langfeldt, 2005). However, this does not seem to be the case here.

Box 2 Measuring Frequency of Prayer

	ISSP 1998	ESS 2002
Question:	About how often do you pray?	Apart from when you are at religious services, how often, if at all, do you pray?
Answers:	1 Never 2 Less than once a year 3 About once or twice a year 4 Several times a year 5 About once a month 6 2-3 times a month 7 Nearly every week 8 Every week 9 Several times a week 10 Once a day 11 Several times a day	1 Every day 2 More than once a week 3 Once a week 4 At least once a month 5 Only on special holy days 6 Less often 7 Never

The third and final indicator of religiosity contained in both surveys taps on the extent to which the respondent believes to be religious. The wordings of the respective questions are very similar, although in the ESS it is stressed that the question is not related to whether the respondent is a member of a religious group or not (cf. Box 3). However, larger differences can be found with regard to the answering scales. Again the polarity of these scales and their degree of differentiation vary. This time the ESS has the finer grained scale in which the numerical values increase with the intensity of the measured attribute. A further difference of the answering scales is that while in the case of the ISSP all answering alternatives are marked verbally the ESS uses an 11-point scale of which only the end points are labeled.

Box 3 Subjective Religiosity

	ISSP 1998	ESS 2002
Question:	Would you describe yourself as ...	Regardless of whether you belong to a particular religion, how religious would you say you are?
Answers:	1 Extremely religious 2 Very religious 3 Somewhat religious 4 Neither religious nor non-religious 5 Somewhat non-religious 6 Very non-religious 7 Extremely non-religious	00 Not at all religious 01 02 ... 08 09 10 Very religious

My expectation is that these three indicators form one common dimension in both datasets. In my view we can interpret this as a validation study. We have a theoretical construct, i.e. religiosity, which we observe through two different devices, i.e. ISSP and ESS, using device specific instruments, i.e. items. Given that the object of interest is sufficiently structured we should be able to make very similar observations with two different, but equally adequate instruments. This is comparable to the observation of an object using different magnifying glasses of different strength. You will see the same item in different degrees of resolution but the relationships between its different parts should be constant or at least almost constant.

As a principal components analysis shows the three indicators load on a common factor; in both studies this factor explains 77% of the total variance (eigenvalue=2.3). With loadings ranging from 0.86 to 0.90 the three indicators equally contribute to the common factor (see Table 4). Thus, the three items which differ slightly between both surveys measure one – and as the subsequent analysis will show – the same dimension, i.e. religiosity, equally well.

Table 4 Principal Components Analysis of Indicators of Religiosity, Factor Loadings

	ISSP-EU 1998	ESS 2002
Attending services	0.87	0.87
Praying	0.90	0.89
Subjective religiosity	0.86	0.88
Common variance	77%	77%

It can be argued that the strength of the relationship between these three indicators of religiosity should differ between religious groups. In other words, the indicators may not be cross-culturally equivalent. For example, for Catholics attending mass on Sunday is a requirement while Protestants can choose where they worship God. To check the extent to which the factorial structure of the three indicators varies with religious group separate principal components analyses were performed.[7] With minor variations the three indica-

7 This approach is advocated by van Deth (1998). Following van Deth's lead Feldkircher (1998) shows that church attendance – one of the indicators used here – is cross-culturally equivalent at least with respect to the five European countries studied by him.

tors indeed measure religiosity well for all groups (see Table 5).[8] Thus, these three indicators measure religiosity equally well for different religious traditions. A similar analysis for the different countries – given in Table A1 in the appendix – shows that these indicators also are equivalent across European countries.

Table 5 Group-specific Results from Principal Components Analysis ESS Data [a]

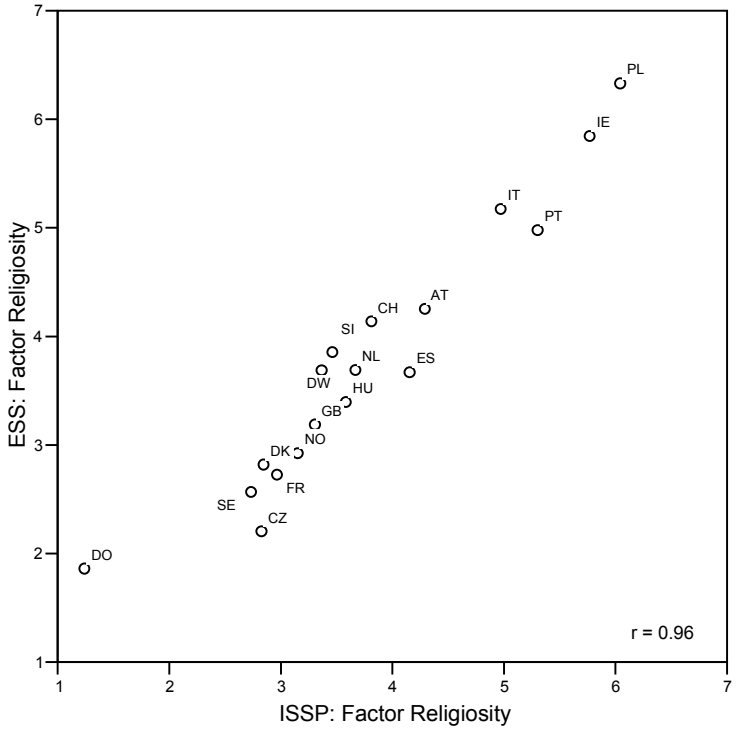
	Highest Eigenvalue	Loadings [b]			Unweighted N
		Religiosity	Attending	Praying	
Catholic	2.1	0.84	0.83	0.85	14,085
Protestant	2.0	0.84	0.79	0.83	6,559
Eastern Orthodox	1.9	0.82	0.78	0.82	2,498
Other Christian	2.3	0.86	0.88	0.90	782
Jews	2.2	0.86	0.85	0.87	1,346
Muslims	1.8	0.78	0.72	0.83	815
Eastern Religions	1.7	0.70	0.68	0.86	125
Other non-Christian	1.6	0.74	0.73	0.71	177
None	1.9	0.80	0.74	0.82	14,802
Total	2.3	0.88	0.87	0.89	41,425

a All ESS countries where included in this analysis.
b Loadings' absolute values are given.

To measure the extent of overall religiosity an index based on the factor scores and rescaled to values between 0 and 10 was constructed, where higher values reflect higher degrees of religiosity. As is clear from Figure 2 ISSP and ESS measure the same trait, the two indexes are correlated on the aggregate level with $r = 0.96$. According to this measure the level of religiosity is particularly low in East-Germany and the Czech Republic. While Italy, Portugal, and especially Ireland and Poland enjoy comparably high levels of religiosity.

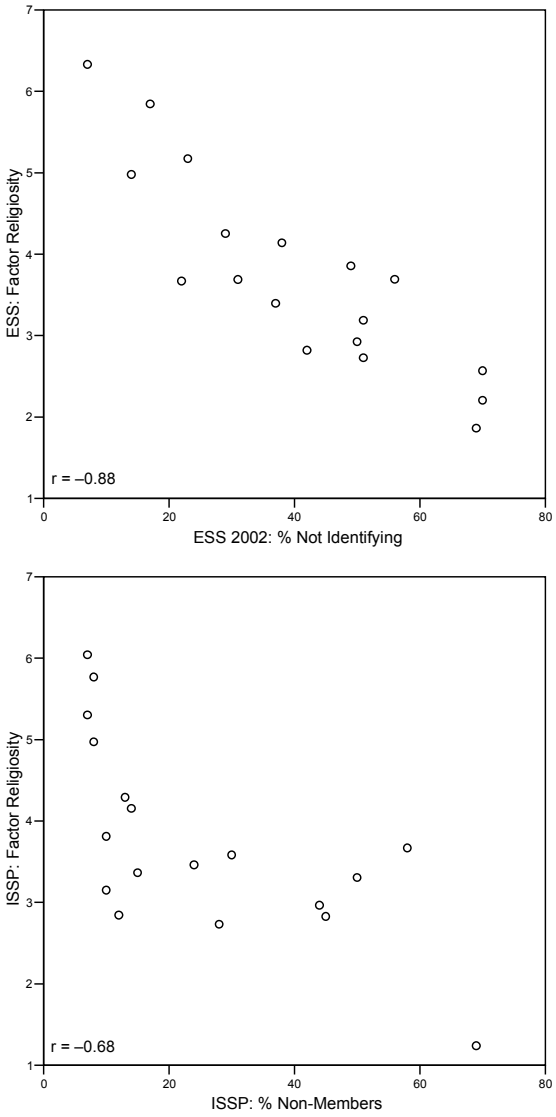
8 It can be argued that the measure of religiosity favors Christian traditions, especially those at home in western Europe. However, the differences are quite small and as was noted above members of non-Christian religious groups are a very small minority in this dataset (and in Europe).

Figure 2 Index of Religiosity in ISSP and ESS



As we saw in the beginning membership rates vary widely between different countries in Europe. The same is true for average religiosity. And Figure 3 shows that there is a strong relationship between membership rates and religiosity on the country level. This statistical relationship in turn can be interpreted as an indication of the validity of the stimulus used to elicit religious affiliation (see Section 2). "Belonging" to a religious group and "Religiosity" seem to measure essentially the same dimension. Thus, "believing without belonging" (Davie, 2002: 5) is rather uncommon in Europe.

Figure 3: Religiosity by Percentage not Identifying with or not being a Member of a Religious Group



4 Conclusion

Religiosity, at least traditional Christian religiosity, can be reliably and validly measured with a set of a few items throughout Europe. The underlying trait seems to be robust and can be captured equally well with slightly different measurement instruments. Good items to capture other forms of religiosity, esoteric, magic or mysticism cross-nationally are still lacking.

As is evident from the results presented above membership in and identifying with a religious group do not have the same meaning in most European countries. Where these measures deviate identification is usually lower, in some instances a lot lower than the membership rate would implicate. The statistical relationship between religiosity and identification with a religious group is stronger than with group membership. Thus, if one measures religiosity and is interested in collecting information one does not capture with this measure the recommendation would be to ask for religious membership. However, this implies that a membership criterion exists and that respondents can indeed judge if they are a member of a religious group or not. In how far this is the case will have to be explored further in the future.

References

Boos-Nünning, U. 1972. *Dimensionen der Religiosität. Zur Operationalisierung und Messung religiöser Einstellungen.* München: Kaiser.

Braun, M., and R. Uher. 2003. "The ISSP and its Approach to Background Variables." Pp. 33-47 in *Advances in Cross-National Comparison. A European Working Book for Demographic and Socio-Economic Variables,* edited by Hoffmeyer-Zlotnik, J. H. P., and C. Wolf. New York: Kluwer Academic/Plenum Publishers.

Bruce, S. 2001. "The social process of secularization." Pp. 249-263 in *The blackwell companion to sociology of religion,* edited by Fenn, R. K. Oxford.

Davie, G. 2002. *Europe: The exceptional case. Parameters of faith in the modern world.* London: Darton, Longman and Todd Ltd.

Davies, N. 1996. *Europe: A History.* Oxford-New York: Oxford University Press.

Demerath., N. J. 2001. "Secularization extended: From religious 'myth' to cultural commonplace." Pp. 211-228 in *The blackwell companion to sociology of religion,* edited by Fenn, R. K. Oxford.

Deth, J. W. van (Ed.). 1998. *Comparative politics. The problem of equivalence.* London: Routledge.

Feldkircher, M. 1998. "Religious orientations and church attendance." Pp. 86-110 in *Comparative Politics. The Problem of Equivalence,* edited by van Deth, J. W. London: Routledge.

Glock, C. Y. 1962. "On the Study of Religious Commitment." *Religious Education (Special Issue)*: 98-110.

Harakas, S. S. 1987. "Christianity in Eastern Europe." Pp. 372-379 in *The encyclopedia of religion,* edited by Eliade, M. New York.

Hill, P. C., and R. W. Hood (Eds.). 1999. *Measures of religiosity.* Birmingham: Religious Education Press.

Jagodzinski, W., and K. Dobbelaere. 1995. "Secularization and Church Religiosity." Pp. 76-119 in *The Impact of Values,* edited by van Deth, J. W., and E. Scarbrough. Oxford: Oxford University Press.

Kecskes, R., and C. Wolf. 1993. "Christliche Religiosität: Konzepte, Indikatoren, Meßinstrumente." *Kölner Zeitschrift für Soziologie und Sozialpsychologie* 45:270-287.

Kecskes, R., and C. Wolf. 1995. "Christliche Religiosität: Dimensionen, Meßinstrumente, Ergebnisse." *Kölner Zeitschrift für Soziologie und Sozialpsychologie* 47:494-515.

Krebs, D., and B. Langfeldt. 2005. *Reliability and Validity of different measures of work orientation: Effects of response scale format on measurement quality.* Applied Statistics 2005, Ribno/Slovenia.

Meulemann, H. 1985. "Die Struktur religiöser Vorstellungen in der Bundesrepublik Deutschland: Eine konfirmatorische Faktorenanalyse." *ZA-Informationen* 16: 40-70.

Partin, H. B. 1987. "Classification of religions." Pp. 527-532 in *The encyclopedia of religion,* edited by Eliade, M. New York.

Pelikan, J. 1987. "Christianity in Western Europe." Pp. 379-387 in *The encyclopedia of religion,* edited by Eliade, M. New York.

Robbers, G. (Ed.). 1996. *Church and State in the European Union.* Baden-Baden: Nomos.

Roof, W. C. 1979. "Concepts and Indicators of Religious Commitment: A Critical Review." Pp. 17-45 in *The Religious Dimension. New Directions in Quantitative Research,* edited by Wuthnow, R. New York: Academic Press.

Schreuder, O. 1991. "Zur Messung der Religiosität in den Niederlanden." *Schweizerische Zeitschrift für Soziologie* 17:484-491.

Slater, W. E., T. W. Hall, and K. J. Edwards. 2001. "Measuring religion and spirituality: Where are we and where are we going?" *Journal of Psychology and Theology* 29:4-21.

Steensland, B., J. Z. Park, M. D. Regnerus, L. D. Robinson, W. B. Wilcox, and R. D. Woodberry. 2000. "The Measure of American Religion: Toward Improving the State of the Art." *Social Forces* 79:291-318.

Voas, D., and A. Crockett. 2005. "Religion in Britain: Neither believing nor belonging." *Sociology* 39(1):11-28.

Wilson, B. R. 1987. "Secularization." Pp. 159-165 in *The encyclopedia of religion,* edited by Eliade, M. New York.

Appendix

Table A1 Countries Included in both the ISSP 1998 and the ESS 2002

	2000 population	Sample sizes [d]		Population weights	
	in million [c]	ISSP-1998	ESS-2002	ISSP-1998	ESS-2002
AU: Austria	8.11	1,002	2,257	0.4274	0.2710
CH: Switzerland	7.19	1,204	2,039	0.3152	0.2659
CZ: Czech Republic	10.27	1,223	1,360	0.4432	0.5695
DK: Denmark	5.34	1,114	1,506	0.2531	0.2674
DO: Germany-East	15.12	1,006	,630	0.7935	1.8108
DW: Germany-West	67.14	1,000	2,289	3.5448	2.2116
ES: Spain	39.47	2,488	1,729	0.8376	1.7215
FR: France	58.89	1,133	1,503	2.7443	2.9548
HU: Hungary	10.02	1,000	1,685	0.5289	0.4485
IE: Ireland	3.79	1,010	2,046	0.1982	0.1397
IT: Italy	57.19	1,009	1,207	2.9929	3.5732
NL: Netherlands	15.93	2,020	2,364	0.4164	0.5082
NO: Norway	4.49	1,532	2,036	0.1547	0.1663
PL: Poland	38.65	1,147	2,110	1.7791	1.3814
PT: Portugal	10.01	1,200	1,511	0.4404	0.4996
SE: Sweden	8.87	1,189	1,999	0.3939	0.3346
SI: Slovenia [a]	1.99	1,006	1,519	0.1044	0.0988
UK: United Kingdom [b]	59.77	1,010	2,052	3.1251	2.1966
Total	422.24	22,293	31,842	1.0000	1.0000

a See http://www.stat.si/doc/pub/rr776-2002/2/T02-01-00.htm (2005/12/02).
b For ISSP data of the Great Britain and Northern Ireland were combined and weighted accordingly.
c Source: Deutschland in Zahlen 2002 (p. 126).
d After weighting with design weights (ISSP: V316; ESS: DWEIGHT).

CONTRIBUTING AUTHORS

BLAIR, JOHNNY
Abt Associates Inc.
Bethesda, Maryland, USA
Email: Johnny_Blair@AbtAssoc.com

BRAUN, MICHAEL
ZUMA (Center for Survey Research and Methodology)
Mannheim, Germany
Email: braun@zuma-mannheim.de

CASPAR, RACHEL
RTI International
Research Triangle Park, North Carolina, USA
Email: caspar@rti.org

CHÁVEZ, NOEL
University of Illinois at Chicago
Chicago, Illinois, USA
Email: nchavez@uic.edu

CHO, YOUNG IK
University of Illinois at Chicago
Chicago, Illinois, USA
Email: youngcho@uic.edu

DEAN, ELIZABETH
RTI International
Research Triangle Park, North Carolina, USA
Email: edean@rti.org

GANZEBOOM, HARRY B.G.
Free University Amsterdam (VUA)
Amsterdam, The Netherlands
Email: hbg.ganzeboom@fsw.vu.nl

HARKNESS, JANET A.
ZUMA (Center for Survey Research and Methodology)
Mannheim, Germany
University of Nebraska-Lincoln
Lincoln, Nebraska, USA
Email: harkness@zuma-mannheim.de

HOFFMEYER-ZLOTNIK, JÜRGEN H.P.
ZUMA (Center for Survey Research and Methodology)
Mannheim, Germany
Email: hoffmeyer-zlotnik@zuma-mannheim.de

HOLBROOK, ALLYSON
University of Illinois at Chicago
Chicago, Illinois, USA
Email: allyson@uic.edu

JOHNSON, TIMOTHY
University of Illinois at Chicago
Chicago, Illinois, USA
Email: timj@uic.edu

KÖRNER, THOMAS
Statistisches Bundesamt der Bundesrepublik Deutschland (Federal Statistical Office)
Wiesbaden, Germany
Email: thomas.koerner@destatis.de

KOLSRUD, KIRSTINE
Social Science Data Services (NSD)
Bergen, Norway
Email: kirstine.kolsrud@nsd.uib.no

LAMBERT, PAUL S.
Stirling University
Stirling, UK
Email: paul.lambert@stir.ac.uk

MCAVINCHEY, GEORGINA
RTI International
Research Triangle Park, North Carolina, USA
Email: mcavinchey@rti.org

MEYER, IRIS
Statistisches Bundesamt der Bundesrepublik Deutschland (Federal Statistical Office)
Wiesbaden, Germany
Email: iris.meyer@destatis.de

MILLER, KRISTEN
National Center for Health Statistics, CDC
Hyattsville, Maryland, USA
Email: KSMiller@cdc.gov
MOSES, LISA
CANFIELD, BETH
EASON, CONNIE

O'ROURKE, DIANE
University of Illinois
Champaign, Illinois, USA
Email: dianeo@srl.uic.edu

PICCININO, LINDA
Abt Associates Inc.
Bethesda, Maryland, USA
Email: Linda_Piccinino@AbtAssoc.com

QUIROZ, ROSANNA
RTI International
Research Triangle Park, North Carolina, USA
Email: quiroz@rti.org

REED, LETICIA
RTI International
Research Triangle Park, North Carolina, USA
Email: lreed@rti.org

ROTHER, NINA
ZUMA (Center for Survey Research and Methodology)
Mannheim, Germany
Email: rother@zuma-mannheim.de

SAGEBIEL, FELIZITAS
University of Wuppertal
Wuppertal, Germany
Email: sagebiel@uni-wuppertal.de

SCHOLZ, EVI
ZUMA (Center for Survey Research and Methodology)
Mannheim, Germany
Email: scholz@zuma-mannheim.de

SKJAK, KNUT KALGRAFF
Social Science Data Services (NSD)
Bergen, Norway
Email: knut.skjak@nsd.uib.no

WARNECKE, RICHARD
University of Illinois at Chicago
Chicago, Illinois, USA
Email: warnecke@uic.edu

WARNER, UWE
CEPS/INSTEAD – Centre d'Etudes de Populations, de Pauvreté et de Politiques Socio-Economiques / International Network for Studies in Technology, Environment, Alternatives, Development
Differdange, Luxembourg
Email: uwe.warner@ceps.lu

WILLIS, GORDON
National Cancer Institute, NIH
Bethesda, Maryland, USA
Email: Willisg@mail.nih.gov

WOLF, CHRISTOF
ZUMA (Center for Survey Research and Methodology)
Mannheim, Germany
Email: wolf@zuma-mannheim.de

ZUCHA, VLASTA
SORA, Institute for Social Research and Analysis
Vienna, Austria
Email: vz@sora.at

ZUMA-Nachrichten Spezial

Die Reihe ZUMA-Nachrichten-Spezial dient dazu, den Forschungsstand größerer Arbeits- oder Forschungsbereiche bei ZUMA zu dokumentieren oder die Ergebnisse von Konferenzen und Symposien vorzustellen (http://www.gesis.org/publikationen/Publikationen/zeitschriften/ZUMA_Nachrichten_spezial/). Bisher sind acht Bände erschienen.

* * *

ZUMA- Nachrichten Spezial Band 1 (vergriffen)

Text Analysis and Computers

Hrsg. von Cornelia Züll, Janet Harkness und Jürgen H.P. Hoffmeyer- Zlotnik

Mannheim, ZUMA, 1996, 132 Seiten, ISBN 3-924220-11-5

Das Heft entstand im Anschluß an eine internationale Tagung zur computerunterstützten Textanalyse, bei der sich Wissenschaftler aus den verschiedensten Disziplinen trafen. Die hier abgedruckten Papiere der eingeladenen Hauptredner dokumentieren den Forschungsstand in vier Bereichen: Computer-Assisted Content Analysis: An Overview *(E. Mergenthaler*); Computer-Aided Qualitative Data Analysis: An Overview (*U. Kelle*); Machine-Readable Text Corpora and the Linguistic Description of Language (*Chr. Mair*); Principle of Content Analysis for Information *Retrieval (J. Krause*). Der Band ist auch als PDF-Datei im Internet verfügbar (http://www.gesis.org/publikationen/zuma_nachrichten_spezial/).

* * *

ZUMA-Nachrichten Spezial Band 2 (vergriffen)

Eurobarometer. Measurement Instruments for Opinions in Europe

Hrsg. von Willem E. Saris und Max Kaase

Mannheim: ZUMA 1997, ISBN 3-924220-12-3

In der Empirischen Sozialforschung finden in Europa Telefoninterviews anstelle von face to face-Interviews zunehmende Verbreitung. Im Rahmen der zweimal jährlich für die Europäische Kommission in Brüssel durchgeführten Repräsentativbefragungen in den Mitgliedsländern der Europäischen Union, den sogenannten Eurobarometern, ergab sich für die Erhebung vom Frühjahr 1994 (EB 41.0) die Möglichkeit, durch eine zeitgleich mit einem weitgehend identischen Fragenprogramm stattfindende Telefonbefragung in den damaligen zwölf Mitgliedsländern der EU, systematisch Effekte der unterschiedlichen Stichprobenansätze und Erhebungsmethoden zu untersuchen. Dabei konnte das Analysespektrum noch durch eine Telefon-Panelkomponente in dreien der zwölf EU-Länder für das face to face-Eurobarometer erweitert werden. Die Beiträge im vorliegenden Buch untersuchen auf dieser Grundlage methodische und methodologische Fragestellungen, die insbesondere für die international

vergleichende Sozialforschung, aber auch für die Markt- und Meinungsforschung in Europa von großer Bedeutung sind. Der Band ist auch als PDF-Datei im Internet verfügbar (http://www.gesis.org/publikationen/zuma_nachrichten_spezial/).

* * *

ZUMA-Nachrichten Spezial Band 3
Cross-Cultural Survey Equivalence.
Hrsg. von J. Harkness
Mannheim: ZUMA 1998, 187 Seiten, ISBN 3-924220-13-1

This volume, the third in the ZUMA-Nachrichten-Spezial series on methodological issues in empirical social science research, is devoted to issues of cross-cultural methodology. The focus is on issues of equivalence, the key requirement in cross-national and cross-cultural comparative research. As the contributions indicate, equivalence is, however, better thought of in terms of equivalencies - in social science surveys and in other standardised instruments of measurement. Contributors come from different countries and continents and from widely differing research backgrounds, ranging from linguistics to survey research and ist methodologies, to cultural anthropology and cross-cultural psychology. They are: Timothy P. Johnson, Fons J.R. van de Vijver, Willem E. Saris, Janet A. Harkness and Alicia Schoua-Glusberg, Michael Braun and Jacqueline Scott, Ingwer Borg: Peter Ph. Mohler, Tom W. Smith and Janet A. Harkness.

* * *

ZUMA-Nachrichten Spezial Band 4 (vergriffen)
Nonresponse in Survey Research
Hrsg. von A. Koch und R. Porst
Mannheim: ZUMA 1998, 354 Seiten, ISBN 3-924220-15-8

This volume, the fourth in the ZUMA-Nachrichten Spezial series on methodological issues in empirical social science research, takes up issues of nonresponse. Nonresponse, that is, the failure to obtain measurements from all targeted members of a survey sample, is a problem which confronts many survey organizations in different parts of the world. The papers in this volume discuss nonresponse from different perspectives: they describe efforts undertaken for individual surveys and procedures employed in different countries to deal with nonresponse, analyses of the role of interviewers, the use of advance letters, incentives, etc. to reduce nonresponse rates, analyses of the correlates and consequences of nonresponse, and descriptions of post-survey statistical adjustments to compensate for nonresponse. All the contributions are based on presentations made at the ‘8th International Workshop on

Household Survey Nonresponse'. The workshop took place in September 1997 in Mannheim, Germany, the home base of the workshop host institute, ZUMA. Twenty-nine papers were presented and discussed, of which twenty-five are included here.

* * *

ZUMA-Nachrichten Spezial Band 5
A review of software for text analysis
Alexa Melina & Cornelia Zuell
Mannheim: ZUMA 1999, 176 Seiten, ISBN 3-924220-16-6

The book reviews a selection of software for computer-assisted text analysis. The primary aim is to provide a detailed account of the spectrum of available text analysis software and catalogue the kinds of support the selected software offers to the user. A related, more general, goal is to record the tendencies both in functionality and technology and identify the areas where more development is needed. For this reason the presented selection of software comprises not only fully developed commercial and research programs, but also prototypes and beta versions. An additional aspect with regards to the kinds of software reviewed is that both qualitative and quantitative-oriented types of research are included. Depending on research purposes and project design the text analyst can profit from available tools independently of their orientation. The following fifteen programs are reviewed: AQUAD, ATLAS.ti, CoAN, Code-A-Text, DICTION, DIMAP-MCCA, HyperRESEARCH, KEDS, NUD*IST, QED, TATOE, TEXTPACK, TextSmart, WinMAXpro, and WordStat and the criteria and methodology used for selecting them are delineated. Der Band ist auch als PDF-Datei im Internet verfügbar (http://www.gesis.org/publikationen/ zuma_nachrichten_spezial/).

* * *

ZUMA-Nachrichten Spezial Band 6
Sozialstrukturanalysen mit dem Mikrozensus
Hrsg. von Paul Lüttinger
Mannheim: ZUMA 1999, 402 Seiten, ISBN 3-924220-17-4

Im Oktober 1998 veranstaltete die Abteilung Mikrodaten von ZUMA die Konferenz "Forschung mit dem Mikrozensus: Analysen zur Sozialstruktur und zum Arbeitsmarkt", an der vorwiegend Nutzer des Mikrozensus teilahmen. Hauptziel dieser ersten Nutzerkonferenz war es, ein Forum für den Informationsaustausch zwischen den Datennutzern und den statistischen Ämtern zu schaffen. Die mehr als 20 Vorträge gingen deutlich über die von den statistischen Ämtern veröffentlichten Standardergebnisse zum Mikrozensus hinaus und sind weitgehend in diesem Band ZUMA-Nachrichten Spezial abgedruckt. Die Autoren sind: Walter Müller; Karl Brenke; Esther Hansch und Michael-Burkhard Piorkowski; Friedhelm

Pfeiffer; Jürgen Schupp, Joachim Frick, Lutz Kaiser und Gert Wagner; Elke Wolf; Dietmar Dathe; Bernd Eggen; Erich Stutzer; Carsten Baumann; Susanne von Below; Thomas Bulmahn; Martin Groß; Reiner H. Dinkel, Marc Luy und Uwe Lebok sowie Wolfgang Strengmannn-Kuhn. Der Band ist als PDF-Datei im Internet verfügbar (http://www.gesis.org/ publikationen/zuma_nachrichten_spezial/).

* * *

ZUMA-Nachrichten Spezial Band 7

Social and Economic Analyses of Consumer Panel Data
Georgios Papastefanou, Peter Schmidt, Axel Börsch-Supan,
Hartmut Lüdtke, Ulrich Oltersdorf (Eds.)
Mannheim: ZUMA 2001; 212 Seiten; CD-Rom

Eine von der Abteilung Einkommen und Verbrauch von ZUMA organisierte Arbeitsgruppe hat sich mit datentechnischem Handling und Analysepotential von komplexen Verbraucherpaneldaten, am Beispiel des ConsumerScan Haushaltspanels der Gesellschaft für Marktforschung (GfK, Nürnberg) beschäftigt und die Ergebnisse in einem Symposium im Oktober 1999 vorgestellt. Die überwiegende Zahl der vorgetragenen Arbeiten, die man als Werkstattberichte ansehen kann, sind in diesem Band abgedruckt. Neben einem detallierten Einblick in die Praxis und das Datenerhebungsprogramm von Verbraucherpanels, wie sie z.B. bei der Marktforschungen der GfK unterhalten werden, enthält der Band z.B. Untersuchungen zu Fragen der Flexibilität von Preisbildungsvorgängen, des Lebensstils im alltäglichen Konsums, der Gesundheitsorientierung im Konsumverhalten, der Umweltorientierung und ihrer Umsetzung im Kauf alltäglicher Haushaltungsprodukte. Der Band enthält eine CD-ROM mit Dokumenten und Codebüchern der aufbereiteten ZUMA-Verbraucherpaneldaten 1995. Der Band ist auch als PDF-Datei im Internet verfügbar (http://www.gesis.org/ publikationen/zuma_nachrichten_spezial/).

* * *

ZUMA-Nachrichten Spezial Band 8

Von Generation zu Generation
Hrsg. von Jan van Deth
Mannheim: ZUMA 2002, 68 Seiten, ISBN 3-924220-23-9

Aus Anlass der Ehrung von Prof. Dr. Max Kaase, Prof. Dr. Walter Müller und Prof. Dr. Hansgert Peisert für ihre langjährige und richtungsweisende Mitarbeit in der Mitgliederversammlung des ZUMA e.V. fand am 14. Juni 2002 eine wissenschaftliche Tagung statt. Der Band enthält Beiträge von Jan van Deth, Hubert Feger, Jürgen Rost, Erwin K. Scheuch, Andreas Diekman und Hans-Dieter Klingemann. Die Beiträge sind auch online verfügbar (http://www.gesis.org/publikationen/zuma_nachrichten_spezial/.)

ZUMA-Nachrichten Spezial Band 9

QUEST 2003

Questionnaire Evaluation Standards

Peter Prüfer, Margrit Rexroth, Floyd Jackson Fowler, Jr. (Eds.)

Mannheim: ZUMA 2004, 216 Seiten, ISBN 3-924220-27-1

This volume, the ninth in the ZUMA-Nachrichten Spezial series on methodological issues in empirical social science research takes up issues of question and questionnaire evaluation. The papers in this volume discuss practical as well as theoretical aspects of questionnaire evaluation. All contributions are based on presentations made at the fourth QUEST (Questionnaire Evaluation Standards) conference which took place from October 21 - 23, 2003 at ZUMA in Mannheim. There were 26 attendees from 9 countries representing 14 organizations: Bureau of Labor Statistics, USA, Center for Survey Research, University of Massachusetts, USA, Institut für Demoskopie Allensbach, Germany, National Center for Health Statistics, USA, National Center for Social Rerearch, U.K., Office of National Statistics, U.K., Statistics Canada, Statistics Finland, Statistics Netherlands, Statistics New Zealand, Statistics Norway, Statistics Sweden, U.S. Census Bureau, ZUMA, Germany. This volume can be downloaded as a PDF file (http://www.gesis.org/publikationen/zuma_nachrichten_spezial/).

* * *

ZUMA-Nachrichten Spezial Band 10

Theory, Laws and Measurement

Michael Braun & Peter Ph. Mohler (Eds.)

Mannheim: ZUMA (forthcoming) ISBN 3-924220-28-X

* * *

ZUMA-Nachrichten Spezial Band 11

Methodological Aspects in Cross-National Research

Jürgen H.P. Hoffmeyer-Zlotnik & Janet A. Harkness (Eds.)

Mannheim: ZUMA 2005, 305 Seiten, ISBN 3-924220-29-8

The idea for this volume was born during the Sixth International Conference on Social Science Methodology in Amsterdam in August 2004, organised by the International Sociological Association Research Committee 33 on Logic and Methodology. Most of the contributions in this volume are proceeding papers from the Amsterdam conference.

The contributions in this volume are organised in four parts. The first part deals with designing and implementing cross-cultural surveys. The second part consists of three papers that deal with different issues of comparability or “equivalence”. The third part of the volume brings together papers on with harmonising socio-demographic information in

different types of surveys. The last section of the volume contains papers that discuss individual socio-demographic variables in cross-national perspective.
This volume can be downloaded as a PDF file (http://www.gesis.org/publikationen/zuma_nachrichten_spezial/).

* * *