

Universität Mannheim
Fakultät für Sozialwissenschaften

Searching for Equivalence
An Exploration of the Potential of Online Probing
with Examples from National Identity

Inauguraldissertation zur Erlangung des akademischen Grades einer Doktorin der
Sozialwissenschaften der Universität Mannheim

Vorgelegt von Katharina Meitinger

Mannheim, 29. Januar 2016

Dekan: Prof. Dr. Michael Diehl
Erstgutachter: Prof. Dr. Michael Braun
Zweitgutachter: Prof. Dr. Eldad Davidov
Drittgutachterin: Prof. Annelies Blom, Ph.D.
Tag der Disputation: 03. Juni 2016

Acknowledgment

Many people have helped, advised, and encouraged me in pursuing and successfully completing this dissertation project. First of all, I would like to thank my first supervisor Michael Braun for providing support, guidance, and continuous discussions. I am grateful for his encouragement, for sharing his vast knowledge on cross-national survey research with me, and for always taking the time for giving me helpful and spot-on feedback. Giving me the opportunity to collect my own data and the time to write on my dissertation, cannot be taken for granted and I am very grateful. I would also like to express my gratitude to Eldad Davidov, my second supervisor, for discussing my initial research ideas, sharing his expertise on measurement invariance tests with me, and his precise and constructive feedback on draft versions of this dissertation. I am also grateful to my supervisors and Annelies Blom for their time and effort as members of the dissertation committee. This dissertation would have been impossible without the support from my other colleagues of the project “Optimizing Probing Procedures for Cross-national Web Surveys.” I enjoyed our wonderful team meetings and I am grateful for their generous feedback on preliminary research results. Among them, most of all, I would like to thank Dorothée Behr, the co-author of my first article, for teaching me the art of writing good articles and for her encouraging words. It was a joy writing the article with her. I am grateful to Lars Kaczmirek for his valuable advice, support, and making me familiar with paradata and regular expressions. I would also like to thank Wolfgang Bandilla for sharing his profound knowledge on web surveys with me.

I wish to thank my former and current colleagues at GESIS–Leibniz Institute for the Social Sciences for encouragement and help, in particular, Evi Scholz, Wanda Otto, Verena Ortmanns, Cornelia Neuert, Kathrin Bogner, Timo Lenzner, and Heike Vester.

Last but not least, I am grateful to my family for their help and my fiancé, Pierre Brice Stahl, for his patience, great advice, and unconditional support throughout the years.

Contents

1. General Introduction	1
1.1 OVERARCHING RESEARCH QUESTION	1
1.2 THE METHOD OF ONLINE PROBING	3
1.2.1 First Goal: Understanding Respondents' Thoughts	4
1.2.2 Second Goal: Uncover Equivalence Problems	6
1.3 ASSESSING EQUIVALENCE	10
1.3.1 Quantitative Approaches	10
1.3.2 Qualitative Approaches	14
1.4 THE RESEARCH FIELD OF NATIONAL IDENTITY	17
1.5 DATA	20
1.6 ANALYZED COUNTRIES	22
1.6.1 Great Britain	22
1.6.2 Spain	23
1.6.3 The U.S.	23
1.6.4 Mexico	24
1.6.5 Germany	24
1.7 OVERVIEW OVER THE FOLLOWING CHAPTERS	25
1.7.1 Comparing Cognitive Interviewing and Online Probing: Do They Find Similar Results?	25
1.7.2 Necessary but Insufficient: Why Measurement Invariance Tests Need Online Probing as a Complementary Tool	26
1.7.3 What Does the General National Pride Item Measure? Insights from Online Probing	27

2. Comparing Cognitive Interviewing and Online Probing:

Do They Find Similar Results?	28
2.1 INTRODUCTION	29
2.2 RESEARCH OBJECTIVES.....	31
2.3 METHODS AND DATA	31
2.3.1 Tested Items.....	31
2.3.2 Methods	32
2.4 RESULTS.....	34
2.4.1 Are There Indications that Response Quality Differs between CI and OP?	34
2.4.2 Do the Methods Uncover Similar Results? The Error Perspective	35
2.4.3 Do The Methods Uncover Similar Themes? The Theme Perspective	39
2.5 DISCUSSION.....	43
2.5.1 Evaluation of Research Questions	43
2.5.2 Researcher Burden in Both Methods.....	44
2.5.3 Limitation	44
2.5.4 Optimal Application	45
References	46
Appendix A. Error Types	49
Appendix B. Screenshots of Probes	50
Appendix C. Key Characteristics of CI and OP Studies	52

3. Necessary but Insufficient: Why Measurement Invariance Tests Need

Online Probing as a Complementary Tool	53
3.1 INTRODUCTION	54
3.1.1 The Quantitative Approach: Tests of Measurement Invariance	55
3.1.2 The Qualitative Approach: Online Probing in a Cross-national Context.....	58
3.1.3 Constructive Patriotism and Nationalism: Theory and Empirical Approaches...	59
3.1.4 Evaluation of Constructive Patriotism and Nationalism	59

3.2 RESEARCH OBJECTIVE	61
3.3 METHODS AND DATA	61
3.3.1 Multi Group Confirmatory Factor Analysis	61
3.3.2 Online Probing.....	62
3.3.3 Comparison of ISSP and Web Survey Results.....	63
3.4 RESULTS OF MEASUREMENT INVARIANCE TESTS: MGCFA	64
3.4.1 Single-country Analysis	64
3.4.2 Measurement Invariance Tests	66
3.5 RESULTS: ONLINE PROBING	68
3.5.1 Category-selection Probe for “Democracy”	68
3.5.2 Specific Probe for “Social Security”	71
3.5.3 Specific Probe for “Fair and Equal”	75
3.6 DISCUSSION.....	78
References	80
Appendix	86

4. What Does the General National Pride Item Measure?:

Insights from Online Probing	87
4.1 INTRODUCTION	88
4.1.1 The General National Pride Item and Its Usage	91
4.1.2 National Identity and Its Elements	94
4.2 ONLINE PROBING AS A METHOD TO UNCOVER RESPONDENTS’ THOUGHTS.....	96
4.3 RESEARCH OBJECTIVES.....	97
4.4 METHODS AND DATA	97
4.5 RESULTS EXPLANATORY FACTOR ANALYSIS	99
4.6 INSIGHTS FROM ONLINE PROBING	100
4.6.1 The Coding Schema	100
4.6.2 Descriptive Results	102

4.6.3 Which Factors Influence the Answer Selection of the GNP Item?	110
4.7 DISCUSSION.....	112
References	115
5. General Conclusion	122
5.1 ARTICLE 1: RESULTS, LIMITATIONS, AND FUTURE RESEARCH.....	122
5.1.1 Results	123
5.1.2 Limitations.....	124
5.1.3 Future research	124
5.2 ARTICLE 2: RESULTS, LIMITATIONS, AND FUTURE RESEARCH.....	127
5.2.1 Results	127
5.2.2 Limitation	128
5.2.3 Future research	129
5.3 ARTICLE 3: RESULTS, LIMITATIONS, AND FUTURE RESEARCH.....	131
5.3.1 Results	131
5.3.2 Limitation	132
5.3.3 Future research.	132
5.4 CONCLUSION	134
References for Introduction and Conclusion	135
Eidestattliche Erklärung.....	151

List of Tables

Table 2.1 Error Types Occurring in Item Battery on Specific National Pride.....	36
Table 2.2 Benefits Mentioned for the Item “Social Security”	41
Table 2.3 Achievements Mentioned for the Item “Arts and Literature”	41
Table 2.4 Social Groups Mentioned for the Item “Fair and Equal”	42
Table S2.1 Variety of Occurring Error Types per Item and Method	49
Table S2.2 Key Characteristics of CI and OP Studies	52
Table 3.1 Items Measuring Nationalism and Constructive Patriotism in ISSP 2013	60
Table 3.2 Comparison of the 2013 ISSP and Web Survey	64
Table 3.3 Single-country: Factor Loadings and Standard Errors	65
Table 3.4 Single-country: RMSEA, CFI, Correlations of Constructs and Errors	66
Table 3.5 MGCFA: Fit Measures of the Measurement Invariance Tests.....	67
Table 3.6 Proportion of Codes for the “Democracy” Item.....	69
Table 3.7 Proportions of Codes for the “Social Security” Item.....	72
Table 3.8 Proportions of Codes for the “Fair and Equal” Item	77
Table S3.1 Single-country (WLSMV): RMSEA, CFI, Correlations of Constructs and Errors.....	86
Table S3.2 Single-country (WLSMV): Factor Loadings and Standard Errors.....	86
Table S3.3 MGCFA (WLSMV): Fit Measures of the Measurement Invariance Test.....	86
Table 4.1 Comparison of the 2013 ISSP and the Web Survey for the GNP Item.....	98
Table 4.2 EFA with Items for Nationalism, Patriotism, and the GNP Item.....	99
Table 4.3 Items Measuring Nationalism, Constructive Patriotism, and GNP in the ISSP	100
Table 4.4 Codes for the Category-selection Probe for the GNP Item in Percent	103
Table 4.5 Percentage of Hidden Patriots and Hidden Nationalists.....	105
Table 4.6 Ordered Logit Regression Analysis of the GNP Item.....	111
Table 4.7 Predicted Probabilities of GNP Values	112

List of Figures

Figure 1.1 The Steps for Establishing Equivalence.....	10
Figure 1.2 The Relation between Concepts, Constructs, Indicators, and Questions	11
Figure S2.1 Screenshot of Closed Item “Scientific and Technological Achievements”	50
Figure S2.2 Screenshot of Category-selection Probe	50
Figure S2.3 Follow-up Probe in Case of a Nonresponse.....	50
Figure S2.4 Screenshot of Closed Item “Fair and Equal”	51
Figure S2.5 Screenshot of Category-selection Probe	51
Figure S2.6 Screenshot of Additional Specific Probe	51
Figure 3.1 Confirmatory Factor Analysis of Nationalism and Constructive Patriotism	62
Figure 3.2 Category-selection Probe for the “Democracy” Item	63
Figure 3.3 Specific Probe for the “Social Security” Item	63
Figure 3.4 Specific Probe for the “Fair and Equal” Item	63
Figure 4.1 The Implementation of the GNP Item in the 2013 ISSP	89
Figure 4.2 Screenshot of Category-selection Probe for the GNP Item.....	98

1. General Introduction

1.1 OVERARCHING RESEARCH QUESTION

Over the last decades, a tremendous increase has occurred in cross-national data production in social science research (Harkness 2008). The large-scale provision and the wide-spread use of cross-national data sets constitute a huge opportunity for the research community but also pose the challenge to develop cross-national comparable survey items (Lynn, Japec, and Lyberg 2006). At the same time, substantive researchers are increasingly aware of the necessity to understand respondents' cognitive processes when answering a survey question (Smith et al. 2011). The recently developed method of online probing can reveal respondents' cognitive processes and helps to assess whether respondents' interpretations of an item differ across countries (Braun et al. 2015).

The overarching goal of this dissertation project is to explore the potential of the method of online probing. Despite its similarities with the method of cognitive interviewing, it remains unclear whether both methods arrive at similar conclusions when applied to the same item. Additionally, it is necessary to study in which research situation priority should be given to cognitive interviewing, in which situation online probing would be preferable, and in which situation a combination of both techniques would be advisable.

In a similar vein, online probing is rather easily applicable for cross-nationally comparative web surveys, which makes it a handy tool for assessing issues of equivalence from a qualitative perspective. Once again, the question remains whether the qualitative approach of online probing and a quantitative approach, such as multigroup confirmatory factor analysis (MG-CFA) (Jöreskog 1971), arrive at similar conclusions when applied to the same constructs. For example, in comparative studies, do they detect the same items as

problematic? More importantly, can online probing explain why some items were flagged as problematic in a quantitative approach? How should these two methods be combined?

Finally, online probing can be used to assess the cross-national comparability of a single-item indicator. Since single-item indicators in comparative research do not allow for quantitative measurement invariance tests, their use has been highly criticized by several researchers (e.g., Ariely and Davidov 2012). Despite this criticism, the use of single-item indicators remains a common practice in cross-national studies (e.g., Heath, Martin, and Spreckelsen 2009; Muñoz 2009; Solt 2011). Contrary to the quantitative approaches, the method of online probing can assess whether a single-item indicator is sufficiently cross-nationally comparable.

To evaluate the potential of online probing in these three areas (comparison with cognitive interviewing, comparison with MGCFAs, and assessment of single-item indicators), it is necessary to find a substantive topic that is relevant but potentially problematic in national and cross-national research. Since the research field of national identity meets both criteria—relevance and problematic nature—it serves as a substantive application for this dissertation.

1.2 THE METHOD OF ONLINE PROBING

Online probing is an innovative method that has been developed recently to assess the validity of survey items. Although research on this method only started in 2010 with the research project “Enhancing the Validity of Intercultural Comparative Surveys,” its feasibility for web surveys has been proven, and design features regarding its optimal implementation have been explored (Behr et al. 2012; Behr et al. 2014b). When applied within a national web survey in Germany, it could shed light on the diverging interpretation patterns of gender items (Behr et al. 2013). Online probing also already has been implemented in several cross-national web surveys to reveal respondents’ diverging or overlapping interpretations and perspectives on survey items assessing xenophobia (Braun, Behr, and Kaczmirek 2013), civil disobedience (Behr et al. 2014a), and satisfaction with democracy (Behr and Braun 2015; for an overview of results regarding online probing, see Braun et al. 2015). Despite its extensive methodological research and substantive applications, online probing, so far, has not been compared with other relevant methods that share similar goals. First, online probing endeavors to open the black box of respondents’ cognitive processes when answering a question and attempts to reveal the different perspectives that respondents adopt when answering a survey question. Second, online probing wants to assess the cross-national comparability of items and uncover problems of equivalence when it is applied to cross-national surveys. With respect to the first goal, online probing stands on the shoulders of cognitive interviewing (Willis 2005). With regard to the second goal, it shares a common aim with the quantitative approach of measurement invariance testing that in most cases applies MGCFA (Jöreskog 1971).

1.2.1 First Goal: Understanding Respondents' Thoughts

By applying the technique of probing, online probing follows the research tradition of cognitive interviewing. Underlying the cognitive interviewing approach is the perspective that respondents carry out a complex cognitive task when they answer survey questions. The response process entails four steps: comprehension, retrieval, judgment, and response (Tourangeau, Rips, and Rasinski 2000). In each step of the response process, errors can occur that can bias the survey results. During the step of comprehension, respondents may have issues with the syntax or they may not understand the vocabulary used. As a consequence, they may not grasp the intended meaning of the question and potentially apply alternative interpretations. In the retrieval phase, a respondent has to access relevant information from memory. Regarding attitudinal questions, respondents must choose between already existing evaluations, vague impressions, general values, and relevant feelings and beliefs (Collins 2015). Depending on the question, some respondents might already have well-formed attitudes and can access pre-existing evaluations, whereas other respondents may not know anything about the question topic and may construct an attitude on the basis of superficial cues present in the survey situation (Tourangeau and Rasinski 1988). During the judgment phase of the response process, respondents form their answer to the survey question. After judgment formation, the respondents still need to decide which answer option most likely corresponds to their opinion. Additionally, they will verify whether their answer selection is socially acceptable, and if not, they may still edit their answer selection (Tourangeau et al. 2000) before responding to the survey question. The optimal situation for survey results occurs when respondents pass through all phases of the response process before selecting an answer category. However, some respondents skip some of the steps of the process, which is called satisficing (Krosnick 1991).

Cognitive interviewing administers “draft survey questions while collecting additional verbal information about survey responses, which is used to evaluate the quality of the

response or to help determine whether the question is generating the information that its author intends” (Beatty and Willis 2007:287). The two dominant variants of cognitive interviewing are think-aloud and verbal probing. When cognitive interviewers apply the *think-aloud* technique, they encourage respondents to verbalize their thoughts while answering a question. In contrast, when applying the *verbal probing* technique, interviewers obtain additional information by asking follow-up questions called *probes* (Beatty and Willis 2007). A different aspect of a question can be targeted by several probe types. For example, a *category-selection* probe inquires about the reasons why a certain answer category has been chosen. With a *specific probe*, a cognitive interviewer can ask for additional information on a particular detail in the question. Finally, *comprehension probes* request a definition of a specific term (Prüfer and Rexroth 2005; Willis 2005). Cognitive interviewing is usually carried out in cognitive laboratories. These face-to-face interviews are typically conducted with a small sample size of 5–15 respondents (Willis 2005) and mostly aim at spotting problematic items during the pretesting phase (Blair and Conrad 2011; Miller et al. 2011).

Online probing: The application of probing techniques in web surveys. The online probing method applies the verbal probing technique used in cognitive interviewing in web surveys. The implementation of this technique within web surveys offers respondents a higher level of anonymity of their answers in comparison to the laboratory situation during cognitive interviewing (Behr and Braun 2015), which potentially reduces social desirability effects in the response process (Bethlehem and Biffignandi 2012). In contrast to cognitive interviewing, online probing can easily realize large samples sizes, which increases the generalizability of the results, enables an evaluation of the prevalence of problems or themes, and can explain the response patterns of specific subpopulations (Braun et al. 2015). Since all probes have to be programmed in advance, all respondents receive the same probe, and the procedure is highly standardized (Braun et al. 2015). Although previous research on open-ended questions has

shown that respondents answer open-ended questions as well or better in web surveys than in paper and pencil surveys (e.g., Holland and Christian 2009; Smyth et al. 2009), online probing studies have reported an elevated number of probe nonresponse and mismatching probe answers (Behr et al. 2014b). This latter finding may be due to a lacking motivating effect of an interviewer. So far, online probing mainly has been used after official data collection to follow-up on problematic items and to assess whether respondents adopt similar perspectives when answering these items (Braun et al. 2015). However, the method also potentially can serve as a pretesting device, or it could be implemented during the actual data collection (Behr et al. 2013). Despite applying the same technique (probing), the methods of online probing and cognitive interviewing seem to have unique strengths and weaknesses. This dissertation follows Braun and colleagues' appeal that the "similarities and differences between the two methods should in any case be further explored" (Braun et al. 2015:195).

1.2.2 Second Goal: Uncover Equivalence Problems

Although online probing is a convenient tool to reveal respondents' thoughts in surveys targeting specific countries, it only unfolds its full potential in cross-national web surveys as a device to uncover equivalence problems.

The goal of revealing equivalence issues addresses the need to test the ever-growing cross-national data volume for its comparability. "Deliberately designed cross-national research has burgeoned in every field that uses survey data, with marked growth in the number, size and diversity of studies undertaken, the disciplines involved, the kinds of instruments used, and the cultures and languages accommodated" (Harkness 2008:57). The number of international surveys is constantly growing with new international surveys, such as the Arab Barometer or the African Barometer that cover areas outside the Western hemisphere (Smith 2010). At the same time, the established cross-national large-scale survey programs, such as the International Social Survey Program (ISSP), the World Values Survey (WVS), and

the European Social Survey (ESS) are incorporating new countries from diverse cultural settings with each new survey round (Smith, Fisher, and Heath 2011). Both developments are indicators of the increasing globalization of surveys (Heath, Fisher, and Smith 2005). Since these surveys make the documentation and data files of their surveys easily available on their websites, the access to cross-national data has been facilitated for researchers (Smith 2010).

The large-scale provision of cross-national data sets constitutes a huge opportunity for research but also a twofold challenge for the data-collecting agency and for researchers using these data sets (Smith 2010) because it adds an additional layer of complexity (Lynn, Japac, and Lyberg 2006) to the creation of data, and on the researcher falls the additional task of assessing the comparability of the used constructs. An important concept in this context is *equivalence*.

The concept of equivalence. The complexity of achieving equivalence becomes apparent when considering the multitude of alternative definitions of equivalence. Already by the end of the 1990s, Johnson (1998) found more than 50 specific terms in his literature review regarding this topic. These various definitions of equivalence also mirror the fact that numerous factors potentially have an impact on the comparability of cross-national data. Although all definitions of equivalence share a reference to the comparability of measured attributes across different populations (Davidov et al. 2014), Johnson distinguishes three topic areas into which most of the definitions fall: (1) *Interpretive equivalence* is “concerned in similarities in how abstract, or latent, concepts are interpreted across cultures” (Johnson 1998:6). This type of equivalence addresses the question whether concepts can be discussed meaningfully within each culture of interest (Hui and Triandis 1985). Closely related to the concept of interpretive equivalence is the “emic-etic” conceptual model (Berry 1969). A concept or behavior is classified as *etic* if it is universal or “understood in a consistent manner across cultural and national boundaries” (Johnson 1998:11). In contrast, *emic* concepts are

culture- or nation-specific, that is, they only are understood by a few cultural groups (Johnson 1998). (2) *Procedural equivalence* is “concerned with the measures and procedures used to make cross-cultural comparisons” (Johnson 1998:7). This aspect of equivalence focusses on the cross-cultural consistency of measurement. (3) *Technical equivalence* is “concerned with the conditions under which surveys are administered” and addresses issues such as the method of data collection. Although this perspective already helps in understanding the multifaceted nature of equivalence, the bias approach from cross-cultural psychology differentiates more precisely between the various nuisance factors that may threaten the comparability of data.

Equivalence and bias. In this psychometric approach, “bias refers to nuisance factors that jeopardize the validity of instruments applied in different cultures. Equivalence refers to the level of comparability of scores across cultures” (He and van de Vijver 2012:3). According to this perspective, a measure is biased when score differences for the indicators of a particular construct do not correspond to the differences in the underlying trait or ability (van de Vijver and Tanzer 2004), which leads to an under- or over-estimation of differences across groups (Davidov et al. 2014). When conducting comparative research, bias needs to be minimized, and equivalence has to be evaluated (He and van de Vijver 2012).

Different types of bias can threaten the equivalence of data. In line with Johnson’s interpretive equivalence, *construct bias* means that the construct measured is not identical across cultures (van de Vijver and Poortinga 1997), whereas a *method bias* “refers to all those biasing effects that are caused by the specific method and context of the measurement” (Fontaine 2005:808). *Samples* might be biased due to cross-cultural variations in sample characteristics (He and van de Vijver 2012), such as issues relating to inconsistent target populations or sampling frames (Heeringa and O’muirheartaigh 2010). The *instrument* might also trigger different response styles across countries (Caramelli and van de Vijver 2013), such as a socially desirable response style, an acquiescent response style, or an extreme

response style (Johnson, Shavitt, and Holbrook 2011). Differences in the procedures or modes of data collection (e.g., face-to-face versus web-mode) can lead to an *administration method bias*. Finally, each item can additionally be affected by an *item bias*. Several factors—such as poor item translation, ambiguous source items, inapplicability of item contents or connotations associated with the item wording in some countries—can trigger an item bias (He and van de Vijver 2012; van de Vijver and Leung 2011).

The typology of biases underlines two facts. First, it is highly complex to achieve equivalent measures. All phases of the survey cycle need to be addressed, from the question development phase to the data analysis phase. Second, the responsibility for minimizing bias and establishing equivalences lies on several actors: the data-collecting agencies, such as the ISSP or the ESS, need to create cross-national data sets by proactively reducing the different types of bias. Additionally, the researchers who apply these cross-national data sets in their studies need to evaluate the equivalence of their constructs before conducting cross-national comparisons. Figure 1.1 is a depiction of an ideal process for establishing equivalence. A detailed discussion of each phase is beyond the scope of this dissertation; instead, it focuses on two methods for assessing the cross-national comparability of constructs and items—MGCFA and online probing.

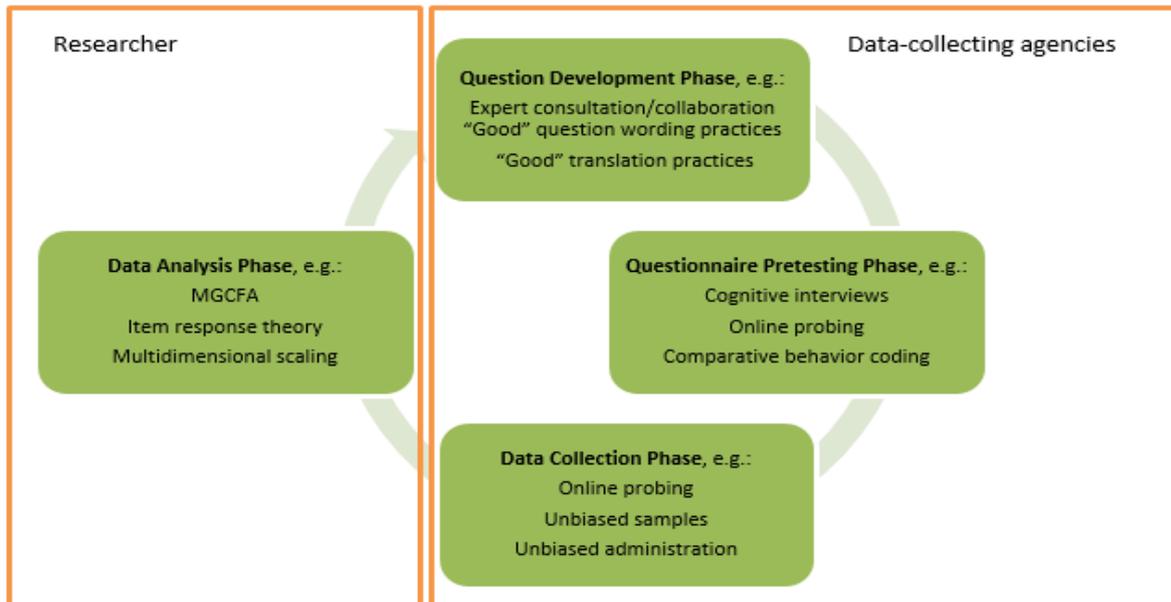


Figure 1.1. The Steps for Establishing Equivalence (Based on Johnson 1998; van de Vijver and Leung 2011; Own Adaptations)

1.3 ASSESSING EQUIVALENCE

The previous discussion made clear that comparability of data should never be assumed; rather, it needs to be assessed before drawing substantive conclusions that are based on cross-national data. Two main approaches to the assessment of equivalence can be distinguished—those using quantitative methods and those using qualitative methods.

1.3.1 Quantitative Approaches

The availability of quantitative approaches to assess the comparability of cross-national data depends on how the intended concepts have been measured. Two research situations can be distinguished: concepts measured with multiple indicators and those measured with single-item indicators.

Assessment of concepts measured with multiple indicators. A variety of quantitative approaches exist that can evaluate the cross-national comparability of data, such as explorative factor analysis (EFA) (Meredith 1964), multigroup confirmatory factor analysis (MGCFA) (Jöreskog 1971), multidimensional scaling (Braun and Scott 1998), multiple

correspondence analysis (MCA) (Blasius and Thiessen 2006), item response theory (IRT) (van de Vijver and Leung 1997), and latent class analysis (LCA) (Kankaraš, Vermunt, and Moors 2011). For an overview of these methods see Braun and Johnson (2010) and Davidov et al. (2014).

The approach of multiple indicator measures draws on the idea that indicators and questions are observable manifestations of indirectly measurable constructs that should represent a theoretical and immeasurable concept. This perspective perceives “survey questions as measurement tools and language vehicles used by researchers to formulate enquiries about indicators chosen to measure specific latent constructs, so as to gain insight into theoretical concepts” (Harkness et al. 2010:41). Although Harkness and colleagues distinguish between indicators and questions, both terms often are used interchangeably in substantive and methodological empirical research. Since usually a distinction is not made in the literature regarding the multiple indicator approach, the present study equates indicators with questions (items).



Figure 1.2. The Relation between Concepts, Constructs, Indicators, and Questions (Harkness et al. 2010:42)

Figure 1.2 illustrates that comparability needs to be established at different levels: “Comparability at the conceptual and indicator levels, however, does not necessarily lead to comparability at the level of constructs. Even if concepts and indicators have the same meaning across different contexts, the link between the two can vary” (Medina et al. 2009:335).

The need to scrutinize these different levels for comparability is addressed by *measurement invariance tests*. If a construct is measured with multiple indicators, this approach enables a researcher to verify “whether or not, under different conditions of

observing and studying phenomena, measurement operations yield measures of the same attribute” (Horn and McArdle 1992:117). Measurement invariance tests help to determine whether differences in measurement scores can be unambiguously interpreted (Horn and McArdle 1992), and they prevent researchers from confusing ambiguous and erroneous data as “real” substantive differences across countries (Steenkamp and Baumgartner 1998). Although a variety of quantitative approaches exist for evaluating measurement invariance (Braun and Johnson 2010), multi-group confirmatory factor analysis (MGCFAs) (Jöreskog 1971) remains the predominant approach that is applied by the majority of substantive researchers (Davidov et al. 2014; van de Vijver 2011). The advantage of this approach is that it can distinguish several levels of invariance that have different consequences for cross-national data analysis. In general, researchers conduct three tests of comparability when applying MGCFAs: configural, metric, and scalar invariance (Meredith 1993; Steenkamp and Baumgartner 1998; Vandenberg and Lance 2000). If *configural* invariance is established, the latent concept can be meaningfully discussed with respect to all countries (Davidov et al. 2014). If *metric* invariance is supported, it is possible to explore cross-national structural relationships, such as regression coefficients. Finally, achieving *scalar* invariance is a precondition for comparing mean values across countries.

Measurement invariance tests are a vital control tool for any researcher who uses data from large-scale cross-national surveys, such as the ISSP or WVS. Since several analysis software tools, such as Amos or Mplus, offer features to conduct measurement invariance tests, these tests are easily accessible to researchers who analyze secondary data. Additionally, this approach can assess the cross-national comparability of a large number of countries. Measurement invariance tests are an ideal tool to detect problematic items and countries that are lacking equivalence but these tests are inadequate for explaining the *reasons* for missing comparability. Although several quantitative approaches—such as the multiple indicators multiple causes model (MIMIC) (Davidov et al. 2014) and the multilevel structural equation

models (MLSEMs) (Davidov et al. 2012)—aim to substantively explain cases of noninvariance, the success of these types of studies heavily hinges on the accuracy of *a priori* hypotheses about cultural differences or the reasons for bias (van de Vijver 2011) that are implemented in the models, and the availability of data to test these hypotheses. As already mentioned, the process of creating equivalent data is highly complex, since numerous factors could have a biasing impact. Additionally, with respect to the known “disturbing” factors, previously unknown and surprising causes might exist. Finding these unexpected causes is one of the major advantages of using qualitative approaches.

Assessment of single-item indicators. The psychometric approach of measurement invariance tests is not necessarily applicable to all research situations because of their reliance on multiple indicators. As Mohler and Johnson (2010) point out, in contrast to psychological research, the situation in social survey research is quite different, since the phenomena of interest often are measured with a limited number of items, or worse, a single-item indicator, an approach that was classified by Johnson as “placing-all-of-the-eggs-in-one basket” (Johnson 1998:23). If the single-item indicator turns out to be invalid, unreliable, or lacking equivalence, the social phenomena that should be measured by this single item is not covered by the survey, since no other item could replace the problematic indicator. Additionally, the assessment of whether a single-item indicator is a valid, reliable, and sufficiently comparable indicator in cross-national research is challenging (Mohler and Johnson 2010), since the usual quantitative approach to test for measurement invariance is inapplicable. A precondition for measurement invariance tests is to measure the concept with multiple indicators (Bollen 1989). Basic data procedures—such as screening answer distributions and the percentage of item-nonresponse across countries or comparing the correlations with benchmark items from a related concept or age across countries—are still possible assessment strategies (Braun and Johnson 2010), although they cannot come close to producing the statistical insights that can

be gained by using multiple indicators. In contrast to measurement invariance tests, qualitative approaches, such as cognitive interviewing and online probing, can assess the cross-national comparability of single-item indicators, and they also can uncover the reasons for missing comparability.

1.3.2 Qualitative Approaches

By 2003, Braun already had cautioned that a purely statistical approach for assessing cross-national data may not suffice, since it also should take into consideration respondents' cognitive processes when answering the items of an international survey:

The array of statistical techniques available to assess the measurement properties of instruments and to detect bias ... is not matched by a similar array of tested procedures to develop equivalent instruments or even a list of tested and reliable guidelines to follow. ... Learning more about how respondents perceive and process items can guide our attempts to improve instrument design for comparative research. (Braun 2003:57)

Two methods address this concern in a cross-national context: cross-cultural cognitive interviewing (CCCI) (Willis 2015) and online probing (Braun et al. 2015). Instead of seeking a statistical evaluation of comparability, both methods want to explain *why* certain items may not work in a cross-national survey. The need to evaluate equivalence from a qualitative perspective increasingly has been recognized by substantive researchers: "As surveys traverse into increasingly diverse territories, such knowledge is perhaps the only way to ensure that survey questions are reliable and valid—not simply for one population of interest, but for all populations of interest" (Smith et al. 2011:492). Additionally, the number of CCCI studies has been growing constantly in recent years (Willis 2015).

However, CCCI studies are being confronted by some challenges that can be solved through an application of online probing (Behr et al. 2013; Behr et al. 2014a; Behr and Braun 2015; Braun et al. 2015). First, the sample size of CCCI studies is usually rather small (e.g., Fitzgerald et al. [2009] report 20 participants). This sample size might be insufficient for achieving saturation of results (e.g., Fujishiro et al. 2010), which means that additional errors

or associations may still be found with a larger sample size (Blair and Conrad 2011). Additionally, the small sample size prevents generalizable conclusions on the differences between country-specific answer patterns (Behr and Braun 2015), although recent CCCI studies have tended to increase sample sizes (Willis 2015). Online probing can achieve large samples sizes through its implementation in web surveys. For example, a previous cross-national online probing study conducted a web survey in six countries with 3,695 respondents (Behr et al. 2014a). Second, online probing circumvents the challenging task of hiring suitable cognitive interviewers. Willis (2015) equates finding such interviewers with a “needle-in-the-haystack requirement” (p. 383) because, optimally, to assess translated questionnaires, interviewers should be at least bilingual, have experience in cognitive interviewing and translations, and be familiar with survey research methods (Liu, Sha, and Park 2013). It also is highly probable that the skill level of interviewers may vary across countries (Gray and Blake 2015). Third, CCCI also involves a harmonizing challenge as house-styles in recruiting respondents and guidelines may differ (Miller et al. 2011). Once again, online probing avoids these harmonization issues. Since each respondent receives the same probe, the procedure is highly standardized.¹

Online probing already has been used successfully to assess several aspects of the cross-national comparability of *multiple-item indicators*. For example, Braun, Behr, and Kaczmirek (2013) assessed whether the word *immigrant*, which is used in each item of an ISSP item battery measuring xenophobia, triggers comparable associations in respondents from different countries. Although respondents from each country thought about different ethnicities, the reported immigrant groups matched the largest and most visible immigrant groups in each respective country rendering the word *immigrant* cross-national comparable in this instance. In a similar vein, Behr and colleagues (2014b) have explained a surprising

¹ However, standardization has been criticized by Willis (2015). Instead of administering standardized probes to respondents, he prefers a flexible approach in which a cognitive interviewer can ask spontaneous probes.

cross-national answer pattern regarding attitudes towards civil disobedience, an item of the ISSP item battery on attitudes towards democracy. They demonstrated that answer behavior was driven by different (miss)understandings of the word *civil disobedience*—some respondents even associated violent acts with the term. Finally, Behr and Braun (2015) evaluated the ISSP *single-item indicator* measuring the satisfaction with the way democracy works. They found that respondents apply a variety of democracy dimensions when answering this item. The fact that respondents from all countries of this survey thought of a multitude of democracy dimensions renders this item cross-national comparable again.

Although the previous online probing results yielded valuable insights into the comparability of survey items, this approach might not be accessible to every researcher, since data needs to be collected. Some researchers may not have the financial resources to conduct a web survey and are limited to using only secondary data files, such as the ISSP or ESS, in their analysis. Due to the large sample size and qualitative nature of the probes, data analysis is more work- and time-intensive than quantitative measurement invariance tests. It is necessary to develop a coding schema, and answers need to be coded and coded a second time to ensure intercoder-reliability. This presupposes some experience in dealing with qualitative data but also the availability of staff that can be involved in the coding process. If several languages are used, probe answers from several countries need to be translated (Behr 2014), which limits the analysis to a small number of survey items and countries that can be tested with this approach. In addition, previous research regarding this topic did not combine online probing with insights from quantitative approaches to assess the cross-national comparability of multiple and single-item indicators. Given the strengths and limitations of each respective method, a combination of both approaches may be a fruitful endeavor.

This perspective also is in line with a growing awareness in comparative survey research that the complexity of creating and assessing cross-national data should be tackled with multiple methodologies (Johnson 1998; Moghaddam, Walker, and Harre 2003; Smith et

al. 2011). Following the idea of data triangulation (Denzin 1970), a mixed method approach combines quantitative with qualitative insights, and also may combine two qualitative or two quantitative methods. If each method arrives at similar results, this convergence is interpreted as consolidating the final research conclusion. Although a successful mixed method approach can compensate for the weaknesses of each method, such studies in cross-national research are still scarce (van de Vijver and Chasiotis 2010).

1.4 THE RESEARCH FIELD OF NATIONAL IDENTITY

National identity is one of the most discussed but least understood concepts of the late 20th century. It is of considerable relevance, with allegiance to state identity, citizenship or ‘nationality’ under threat not only from the rise of different national identities within states, but also by the growth of systems (such as the European Union) that seek to encompass a plurality of states (McCrone 1998:1).

National identity is a *relevant research field*, since national identity and national pride are important parts of personal identity in contemporary societies (Haller 2009). National identity is the “positive, subjectively important emotional bond with a nation” (Tajfel and Turner 1986) and also “the cohesive force that holds nations together and shapes their relationships with the family of nations” (Smith and Jarkko 2001:1). Miller and Ali (2014) have classified national identity as the cement or glue that enables modern and culturally diverse societies to function effectively.

Several factors potentially impact national identity, which makes it an interesting social phenomenon to study. With the growing impact of globalization that has led to an increasing economic and political integration between states, a theoretical interest in the role of national identity in this process has increased (Sinnott 2006). At the same time, some nation states are profoundly changing by “shifting borders and [they] develop over time due to wars, revolution or other political change” (MacInnes 2006:104). New nation states are founded, while others fall apart or “unite themselves with neighboring states into large macro-

regional associations (European Union)” (Haller 2009:172). Other nation states are subject to the internal pressures of regional separatist movements’ growing strengths (Medrano and Gutiérrez 2001). Additionally, increasing cross-frontier migration renders the borders of nation states more and more “porous” (MacInnes 2006).

The *empirical* approach to measuring the different aspects of national identity can be roughly divided into two main research streams. On the one hand, researchers differentiate between a *civic* and an *ethnic* form of national identity (e.g., Reeskens and Wright 2013; Smith 1991). This approach is concerned with the criteria that are applied by respondents to differentiate between who is perceived as a fellow citizen and who is not. On the other hand, some researchers perceive *patriotism* and *nationalism* as two sub-dimensions of national identity (Adorno et al. 1950; Blank and Schmidt 2003; Davidov 2009, 2011; Kosterman and Feshbach 1989; Schatz, Staub, and Lavine 1999). Patriotism and nationalism are two expressions of national affection (Latcheva, 2011). This dissertation focuses on the second research tradition that studies national affections such as patriotism, nationalism, and national pride.

The measurement of national identity also is *problematic* for two reasons. First, the research field has struggled to find a *consensus on the definitions and conceptualization* of the different elements of national identity (Davidov 2009; Latcheva 2011). Similar to many other research fields, an overwhelming number of definitions of specific social elements of national identity exist. Additionally, the elements of national identity are measured with different items, which raises doubts about the comparability of research results. Moreover, the same items serve as measures for different and partly contradictory concepts. A prime example of measuring various concepts with the same indicator is the use of the general national pride item (e.g., “How proud are you of being German?”) that was employed as an indicator for various concepts, such as nationalism (Solt 2011), patriotism (Ariely 2012), and national attachment (Elkins and Sides 2006).

Second, several researchers have pointed to *measurement issues and problems of data quality* in regard to measures of national identity. For example, some researchers working with theories of national identity seem reluctant to use the available data sets on this topic:

Given the availability of these data and their potential relevance, one must wonder whether the relative paucity of research is due to weaknesses in the way in which national and other levels of identity have been operationalized in the surveys in question. Could it even be that concepts such as identity are impossible to capture in mass survey research? (Sinnott 2006:211)

Although Sinnott's argument about the paucity of empirical research using this data does not hold any more—given that the current ISSP bibliography for the Module on National Identity already lists 512 entries (Smith and Schapiro 2015)—several problematic issues regarding the empirical data have been reported. For example, the current large-scale comparative surveys have been criticized for using suboptimal scales to measure national identity (Sinnott 2006). Heath and colleagues remarked that “[t]he huge variety in sampling methods, modes of administration, response rates, and response biases ... leave[s] plenty of scope for the creation of artefactual results” (p. 303). Additionally, they found errors of observations such as acquiescence bias, social acceptability bias, and the use of extreme response categories. Even more troublesome, they detected a lack of equivalence of meaning for some items in the item battery intended to measure the criteria of national belonging. Concerns regarding the ambiguity of items also have been raised regarding the ISSP item battery of domain specific national pride. By 2001, Smith and Jarkko already had criticized the item “pride in the fair and equal treatment of all groups in society” that is part of this item battery:

Rankings are somewhat hard to interpret because the item may be understood in different ways across countries and individuals. It may be thought of as referring to ethnic, racial, and religious groups in some countries and to class and income groups in others. (Smith and Jarkko 2001:7)

Further items of this item battery were the target of the same critique, such as “pride in the country's history,” “pride in the country's political influence,” “pride in science and technology,” and “pride in the military” (Bonikowski 2009; Hjern 1998). Additionally,

previous qualitative research has revealed several reasons why these items are error prone in regard to the Austrian context (Fleiß et al. 2009; Latcheva 2011), such as a too broad and unspecific formulation of time spans, key terms that allow respondents to adopt various perspectives when answering this item, an insufficient number of answer categories, and context effects in the questionnaire (Latcheva 2011).

Since this dissertation sets out to compare the methods of online probing, cognitive interviewing, and MGCFA in regard to their potential to detect problematic issues at the item level, the field of national identity seems to be a fitting substantive application for method comparisons.

1.5 DATA

This dissertation uses data from the 2013 ISSP module on National Identity. Originally, the ISSP was established in 1984 by Australia, Germany, Great Britain, and the United States (Skjåk 2010), and currently, it has 45 member countries (ISSP 2015a). It is a continuing annual program that runs as an add-on to respective national surveys (Smith 2009). The ISSP questionnaires address issues that are highly relevant for social science research, such as “role of government,” “social inequality,” and “family and changing gender roles.” The modules often are repeated in part after several years, thus creating a research design that combines a cross-cultural perspective with a cross-time perspective (Skjåk 2010). As the ISSP bibliography with its 5,700 entries proves, the ISSP data is widely used in social science research (ISSP 2015b). The ISSP searches the input of different country experts at an early stage of its questionnaire development by using a multi-cultural drafting group (Harkness 2008). It also decides on themes for future modules in plenary meetings:

That distinguishes ISSP from “imperialistic” forms of organized research—where one national team figures out a study and implements it in foreign countries relying at best on some technical advice from indigenous pollster, only—and makes the most efficient use of the competences of the national team. (Braun and Uher 2003:35)

Additionally, the ISSP has implemented a Methodological Committee that is assisted by the Methods Working Groups (Skjåk 2010) that works on issues of translation, demographic comparability, and questionnaire design (Medina et al. 2009). However, since the ISSP is an add-on to national surveys, it also is bound to a certain degree to national traditions, which hinders the homogenization process of sampling, the mode of administration, and questionnaire construction (Braun and Uher 2003; Skjåk 2010).

The ISSP Module on National Identity assesses “nationalism and patriotism, localism and globalism, and diversity and immigration” (Smith 2009:9) in a cross-national context. Fielded for the first time in 1995, the module was partially replicated in 2003 and in 2013. The third Module on National Identity was administered in 33 countries with 45,297 valid respondents in a fielding period from 2012 to 2015, with most of the countries completing their data collection in 2013 and 2014 (GESIS 2015). A limited data set of five countries was used for the current study. These countries include Germany ($N=1,717$), Great Britain ($N=904$), the U.S. ($N=1,274$), Mexico ($N=1,062$), and Spain ($N=1,225$), which are the countries that also were targeted in the other data sources that were used in this dissertation.

Additionally, data was collected within the DFG funded project “Optimizing Probing Procedures for Cross-national Web Surveys” that ran from 2013 till 2016. The project’s goal was to improve the recently developed method of online probing. In two of the web surveys that were conducted within this project, items from the ISSP Module on National Identity were replicated. The first web survey was conducted in September 2013 in Germany. The 532 respondents were drawn from a non-probability online panel with quotas for age (18–30, 31–50, and 51–65), gender, and education (lower and higher). Alongside this web survey, the same ISSP items were tested in the GESIS Pretest Lab with 20 cognitive interview participants. The German web survey and the transcriptions of the cognitive interviews were the data basis for the method comparison of online probing and cognitive interviewing in Chapter 2 of this dissertation. The second web survey was conducted with 2,685 participants

in May 2014 in Germany, Great Britain, Mexico, Spain, and the U.S. Similar to the German web survey, participants were drawn from a non-probability online panel with quotas for age (18–30, 31–50, and 51–65), gender, and education (lower and higher). In both web surveys, several items were followed by probes. To avoid respondents' frustration, the maximum number of probes for each respondent was set at nine. The probe responses of the second web survey served as the data basis for the method comparison between online probing and MGCFA (Chapter 3 of this dissertation) and the assessment of online probing as an evaluation tool for single-item indicators (Chapter 4 of this dissertation).

1.6 ANALYZED COUNTRIES

Since the data was collected within a research project, the countries in which the web survey was conducted were preset. The selection of the five countries (Germany, Great Britain, Mexico, Spain, and the U.S.) originally was based on considerations, amongst others, about language. Although the five countries in this study were not deliberately chosen, they nevertheless provide an interesting variety for the assessment of measures of national identity.

1.6.1 Great Britain

Although some commonly shared symbols of British culture exist as a source of national pride, such as British democracy, the British sense of fair play (Bechhofer and McCrone 2013), and the welfare state (Tilley and Heath 2007), “British identity shows a general pattern of fragmentation” (Cohen 1994:35). On the one hand, Great Britain has lost some of its key sources of identification in the process of modernization, such as the end of the Empire, the loss of its global influence, and the declining importance of Protestantism (Tilley and Heath 2007). On the other hand, Great Britain is a multicultural and multinational state. Postwar migration from its previous Empire (McCrone 2002) and current immigration have created an

ethnic pluralization of British society (Tilley and Heath 2007). In addition, the rise of Scottish and Welsh nationalism has increased the importance of “territorial identities,” a process that also is gaining in significance for English nationalism (Bechhofer and McCrone 2010).

1.6.2 Spain

Territorial identities also are present in Spain. Due to the strong nationalisms of Catalonia and the Basque Country (Medrano and Gutiérrez 2001), Bollen and Medrano (1998) even speak of Spain as an example of incomplete nation building. Spain also is a relatively young democracy, since its constitution only dates back to 1978. In the years following Franco’s authoritarian regime, the newly established constitutional monarchy needed to redefine Spain’s national identity, which led to a high degree of decentralization and a decline of the importance of Catholicism for the Spanish national identity (Muñoz 2009). After a long period of uninterrupted economic growth, in recent years, Spain has been hit by a major economic crisis (Encarnación 2009). In addition to the tremendous increase in the country’s unemployment rate, the economic crisis in Spain has led to an increased level of political distrust due to a negative perception of the political responsiveness of representative institutions and an increasing perception of political corruption (Torcal 2014).

1.6.3 The U.S.

The U.S. is an influential military, economic, and cultural superpower (Hutcheson et al. 2004), which is reflected in the extremely high general national pride levels that additionally increased in the wake of the terrorist attacks in 2001 (Smith and Kim 2006). According to Schildkraut (2014), “American identity is rooted in a complex, and often contradictory, set of beliefs that includes, but is hardly limited to, the liberal creedal tradition of individualism, minimal government intervention into private life, hard work, equal opportunity, and political freedom” (p. 447). In contrast to Great Britain and Spain, religion continues to be of

importance to American identity, with many Americans considering the U.S. as a Christian nation (Merino 2010). The U.S. also has a high level of immigration-related diversity, and racial and ethnic differences are important factors in the perception of national identity (Schildkraut 2014).

1.6.4 Mexico

In contrast to the U.S. and Great Britain, Mexico is not a country of immigration; rather it is a country of emigration with many of its citizens currently leaving for the U.S. (Theiss-Morse and Wals 2014). Given the discrepancy between the economic power of the U.S. and Mexico, the U.S. serves as the “predominant other” for Mexican national identity, with pride in its past Indian civilization and the Mexican revolution serving as other important features constituting Mexican national pride (Morris 1999). In recent years, incidences of violence and criminality have increased in Mexico. For example, between 2006 and 2011, Mexico registered 47,515 crime-related deaths (Gonzalez 2012). Additionally, the democratic situation in the country has deteriorated due to the high levels of corruption in the political system, which was described in the 2014 Annual Report of Freedom House in which Mexico was downgraded from a “free” to a “partly free” democracy (Freedom House 2015).

1.6.5 Germany

Germany is a powerful economy with a highly developed social security system and a stable democracy (Freedom House 2015). However, its national identity is closely intertwined with its history. In the aftermath of WWII and the Nazi regime, a public narrative was established in West Germany that prohibited the open expression of national pride. “National pride” was seen as connected to racism and chauvinistic attitudes, and the expression of national pride still triggers right-wing connotations to this day (Miller-Idriss 2009). This pride taboo translates to consistently low levels of national pride in cross-national surveys (Kelley and

Evans 2002) because German respondents feel it is not permissible to overtly express national pride, a phenomenon that Smith and Jarkko (2001) have called the “war guilt effect.”

1.7 OVERVIEW OVER THE FOLLOWING CHAPTERS

1.7.1 Comparing Cognitive Interviewing and Online Probing: Do They Find Similar Results?

The first article (Chapter 2, the article is joint work with Dorothee Behr) evaluates the potential of online probing vis-à-vis another qualitative method by analyzing the similarities and differences of the methods of online probing and cognitive interviewing. Since online probing builds on the theoretical frameworks of cognitive interviewing and applies the same technique—namely probing—as its predecessor, an overlap in research results might be expected. However, both methods have distinctive features that could produce a variation of research results. These two methods are conducted in different modes, have diverging sample sizes, vary in their level of interactivity, and differ in their typical research goals. Thus, the article aims to answer the following research questions:

1. Are there indications that response quality differs between CI and OP?
2. Do cognitive interviewing and online probing methods produce similar results? How do sample size and the divergent levels of interactivity affect the performance of these two qualitative methods?

The similarity of results is assessed from two perspectives: an *error perspective* that evaluates the capacity of both methods to detect problems in question wording, and a *theme perspective* that compares both methods in regard to the content and variety of the associations that are mentioned by respondents of these methods.

Since this article strived to evaluate the potential of both methods in error detection, an item battery was chosen that previous research already had flagged as problematic. Several past studies had indicated problematic issues concerning the item battery on domain-specific

national pride in the ISSP Module on National Identity (Fleiß et al. 2009; Latcheva 2011; Smith and Jarkko 2001), and so, this item battery was chosen for this method comparison.

1.7.2 Necessary but Insufficient: Why Measurement Invariance Tests Need Online Probing as a Complementary Tool

Another important goal of online probing is to evaluate the cross-national comparability of items, a research aim that also is pursued by quantitative psychometric approaches, such as MGCFA. An evaluation of the potential of online probing vis-à-vis this quantitative approach is the topic of the second article (Chapter 3) of this dissertation. The following research questions are addressed:

1. Do online probing and MGCFA arrive at similar conclusions?
2. Can the insights from online probing help to explain instances of missing comparability?
3. What is the optimal way to combine both methods in different situations?

To compare online probing and MGCFA, it is necessary to apply them to constructs measured with multiple items, which is a precondition for the application of quantitative measurement invariance tests. The constructs of constructive patriotism and nationalism offer an intriguing substantive application for the method comparison. On the one hand, the results from the first article of this dissertation project showed that several items measuring the construct of constructive patriotism were problematic with respect to the German context. On the other hand, Davidov (2009), using MGCFA to develop and test measures of constructive patriotism and nationalism for measurement invariance, established metric invariance of these constructs. In contrast, scalar invariance tests failed for these constructs. This enables a comparison of structural relationships, such as regression coefficients, but prohibits a cross-national comparison of latent means.

1.7.3 What Does the General National Pride Item Measure? Insights from Online Probing

Finally, the third article (Chapter 4) critically assesses the appropriateness of using single-item indicators for cross-national studies with online probing. Since quantitative approaches to measurement invariance testing need multiple item indicators to evaluate the cross-national comparability of a construct (Bollen 1989), methods like MGCFA cannot be used to test constructs that are measured with one indicator. In contrast, online probing can be used in this context as a device to reveal equivalence problems.

The general national pride item serves as an interesting substantive application. It is a popular item in cross-national studies of national identity. Despite it being a single-item indicator, it has been used to measure a variety of concepts related to national identity, such as national attachment (Elkins and Sides 2006), nationalism (Solt 2011), and patriotism (Ariely 2012). At the same time, it is highly probable that respondents' answer selections for this item might be distorted due to several issues, such as social desirability effects. The third article addresses the following research goals:

1. It assesses the suitability of the general national pride item as a cross-national indicator for the different elements of national identity, and explores the full variety of respondents' associations when answering this item.
2. It assesses the prevalence of potentially problematic issues.
3. It demonstrates how exploratory factor analysis, online probing, and regression analysis can be combined to detect and explain equivalence issues and to assess the distorting impact of revealed problems.

The results of the three articles, the limitations of the three studies, and a critical outlook are discussed in an overall conclusion (Chapter 5).

2. Comparing Cognitive Interviewing and Online Probing: Do They Find Similar Results?²³

Abstract

This study compares the application of probing techniques in cognitive interviewing and online probing. Even though the probing is similar, the methods differ regarding typical mode setting, sample size, level of interactivity, and goals. We analyzed probing answers to the ISSP item battery on specific national pride. While probing answers in cognitive interviewing show indications for a higher response quality, online probing can compensate through a larger sample size. Therefore, both methods have complementary strengths with regard to error detection and themes.

² A joint work with Dorothée Behr.

³ Preliminary results were presented at the International Workshop on Comparative Survey Design and Implementation (CSDI), March 27-29, 2014, Bethesda, USA, and at the XVIII ISA World Congress of Sociology, July 13-19, 2014, Yokohama, Japan.

2.1 INTRODUCTION

Cognitive interviewing (CI) and online probing (OP) both aim to reveal the cognitive processes respondents use when answering survey questions. Beatty and Willis (2007) defined CI as “the administration of draft survey questions while collecting additional verbal information about survey responses, which is used to evaluate the quality of the response or to help determine whether the question is generating the information that its author intends” (p. 287). CI usually comes in two dominant variants: think-aloud and verbal probing. Think-aloud encourages respondents to verbalize their thoughts while answering a question. In verbal probing, the interviewer obtains additional information by asking follow-up questions called probes (Beatty and Willis 2007).

Recently, OP has been developed as a complementary method. It implements probing techniques within web surveys (Braun et al. 2015; Murphy et al. 2013). Although the probing techniques are similar, typical implementations of CI and OP vary in several ways: mode setting, sample size, level of interactivity, and goals.

Both methods differ in the *mode setting* and its differential impact on respondent motivation and degree of anonymity. CI is usually conducted in a face-to-face (laboratory) setting and involves an interviewer who can motivate the respondent. In web surveys, a motivating interviewer effect is missing, which could ease satisficing (Krosnick 1991). Previous OP studies thus report elevated probe nonresponse and mismatching probe answers (Behr et al. 2014).

Closely connected to the mode setting is the divergent *sample size*. The traditional CI procedure involves small sample sizes of five–15 respondents with ideally iterative testing rounds (Willis 2005), although it is possible to conduct CI with a substantially larger sample size (e.g., 100 participants). While small sample sizes allow for in-depth interviews with respondents, they risk missing potential errors (Blair and Conrad 2011); moreover, they increase the probability that identified errors are false positives (Conrad and Blair 2009). Due

to its implementation within web surveys, OP can realize large sample sizes. These allow for a quantitative data analysis and for judging the prevalence of an error or theme; furthermore, they can help avoid false positives (Behr et al. 2014).

CI and OP also vary in their *level of interactivity*. Cognitive interviewers can ask follow-up probes in case of insufficient answers. These *emergent* probes can be directly adapted to the original probe response. OP studies have to foresee any probes or potential problems, such as nonresponse. Hence, researchers are restricted to *conditional* follow-up probes in OP (i.e., probes that are conceived and implemented prior to the fielding) (see Willis [2005] on terminology). Nonetheless, the missing interactivity in OP increases standardization because every participant receives identical stimuli. Conrad and Blair (2009) found that proactive probing during CI threatens the comparability of results between different interviewers. OP prevents such potential interviewer effects. Interactivity also applies to respondents: CI respondents can make spontaneous remarks prior to any probe and thus provide voluntary feedback. Spontaneous comments are impossible in OP.

Finally, researchers are currently using these methods for slightly different *goals* and at different *stages of the data collection process*. Most researchers see the main purpose of CI in error detection during the pretesting phase, which we call *error perspective* here (e.g., Blair and Conrad 2011). OP has so far mainly been used *after* official data collection to follow-up on problematic items and to assess whether respondents think of similar themes when answering these items (Braun et al. 2015). OP has therefore applied a *theme perspective*.

2.2 RESEARCH OBJECTIVES

In this article, we aim to answer the following research questions:

1. Are there indications that response quality differs between CI and OP?

We gauge response quality by looking at probe nonresponse and length of responses.

2. Do CI and OP methods produce similar results? And how do sample size and the divergent level of interactivity affect the performance?

After assessing the response quality, we compare CI and OP from two different perspectives. First, we adopt an error perspective to assess the capacity of both methods to detect problems in question wording. Second, we compare CI and OP from a theme perspective. In both perspectives, we first look at the number of errors/themes detected, then we compare whether both methods uncover similar errors/themes, and finally we assess the impact of interactivity.

2.3 METHODS AND DATA

2.3.1 Tested Items

The tested items come from the 2003 ISSP module on national identity. Ten items from the item battery on specific national pride were probed. The closed item was “How proud are you of Germany in each of the following?” Respondents then had to rate 10 different domains on a four-point scale running from *very proud* to *not proud at all*. A “don’t know” option was given. The items were:

- A. the way democracy works
- B. its political influence in the world
- C. Germany’s economic achievements
- D. its social security system

- E. its scientific and technological achievements
- F. its achievements in sports
- G. its achievements in the arts and literature
- H. Germany's armed forces
- I. its history
- J. its fair and equal treatment of all groups in society

The item battery was selected because previous research has identified several problems. In their quantitative analysis, Smith and Jarkko (2001) pointed to potential problems in the cross-national context. For example, they suggested that the item “fair and equal” could be understood differently across countries because it might refer to different social groups. Additionally, CI studies have been undertaken, notably in Austria: Fleiß and colleagues (2009) criticized the vague formulation of key terms such as “democracy,” which allowed respondents to adapt multiple perspectives when answering. In her CI study, Latcheva (2011) pointed to issues such as unspecified historical time spans for several items (e.g., history) and to respondents' problems with the word “pride.” Given the wide-ranging criticism, we regard this item battery as a suitable instrument for comparing CI and OP: Several error types and different interpretation patterns can potentially be detected.

2.3.2 Methods

The overall goal of this study is to compare probing in CI and OP when it is conducted “as usual.” The study compares data from cognitive interviews conducted with 20 German respondents in April 2013 with a web survey conducted with 532 German respondents in September 2013. Web survey participants were drawn from a national non-probability online panel, and CI respondents were locally recruited in the area of Mannheim. Three researcher and two student assistants conducted the cognitive interviews. All interviewers had previous experience in conducting CI, received specific CI training, and participated in test interviews.

Both studies used quotas for age (18–30, 31–50, and 51–65), gender, and education (lower and higher education). The item battery in both methods was part of longer questionnaires containing questions about national identity, gender roles, and political participation. While CI was completed by all respondents, the break off-rate for the web survey was 11.5 percent.

All CI participants received probes after *each* item of the item battery. The cognitive interviews were conducted following the GESIS “house style,” which included an introduction to encourage respondents to point out potential flaws in questions. CI respondents received 15 anticipated, standardized probes at the item battery (10 category-selection probes and five specific probes). During CI, interviewers could follow-up on anticipated probes with emergent probing (see Willis [2005] on terminology). Spontaneous remarks from participants prior to any probe were recorded.

Due to the risk of increased break-off rates in OP, we refrained from probing the entire battery with each respondent. Instead, we divided the sample by five and thus obtained groups of 105–110 respondents. Every web respondent answered each closed item of the battery (administered on separate screens); for a randomly selected set of two items, each group then received three probes (on separate screens). The first item was followed by a category-selection probe asking why a certain answer category had been chosen (Prüfer and Rexroth 2005). The second item was followed by both a category-selection probe and a specific probe asking for additional information on a detail of the question (Willis 2005). The survey software could detect cases of probe nonresponse. When a respondent gave a probe nonresponse answer, a conditional probe was triggered with a motivational sentence (e.g., “Please consider the question again. Your answer is very important for this research project.”).

The online section of this article contains screenshots of the probes. A limitation of this study is that the web survey was not clearly framed as a pretest study and did not encourage respondents to spot errors; this would have been counterproductive to the main

research goals of the web survey. Instead, respondents were encouraged to express their opinion when answering the probes.

2.4 RESULTS

2.4.1 Are There Indications that Response Quality Differs between CI and OP?

We gauged the response quality by evaluating the percentage of probe nonresponse and response length. We differentiated three types of probe nonresponse: (1) Respondents who lacked information or knowledge (DK answers, e.g., “I don’t know”); (2) respondents who gave incomplete and uninterpretable answers (non-substantive responses, e.g., “It is my feeling,” “Just like that”); and (3) respondents who refused to answer (refusals, e.g., “Don’t want to answer,” empty text boxes). CI probes had almost no probe nonresponse: DK probe answers and non-substantive probe responses occurred rarely (0–1 respondent per probe/item), and none of the CI respondents refused to answer.

In contrast, nonresponse per probe varied between 10.5 percent and 31.4 percent in OP: DK probe answers fluctuated between 0 and 7.6 percent. Additionally, OP was affected by non-substantive responses for some items (up to 4.7 percent). Furthermore, between 7.6 percent and 17.3 percent of OP respondents refused to answer. The increased probe refusals especially support the assumption that interviewers can improve the quality of probe responses.

Regarding answer length, we found that CI answers were longer than OP responses. Taking into account only substantive answers to anticipated probes, CI participants answered a probe with 47 words on average; OP respondents wrote 12 words on average.

2.4.2 Do the Methods Uncover Similar Results? The Error Perspective

Our first step was to compare both methods from an error perspective. For this purpose, we applied an error coding schema that built loosely on a schema from DeMaio and Landreth (2004). The schema ordered errors along Tourangeau's distinction of the components of the response process, namely comprehension, retrieval, judgment, and response (Tourangeau et al. 2000). This schema served as a starting point to analyze the probe answers. However, in an iterative process, it was modified to fit the data of the study. The full coding schema is available from the authors on request.

Responses can be classified into different error types. Table 2.1 shows an overview of error types that applied to the entire item battery and their occurrence in both methods (CI: 20 interviews, OP: 105–110 respondents). The items were coded by one researcher, but were coded a second time by student assistants. The intercoder agreement varied between 91 percent and 100 percent for the different items/probes. The coding team discussed rare incidences of mismatching codings and made the final decision about the appropriate codings.

How many errors are identified? During CI, 95 error incidences were found in the entire item battery; during OP, 112 error incidences were found in total. Taking into account the different sample sizes, CI respondents provided, on average, more indications for errors per item and probe ($M=.32$, $SD=.23$) than their OP counterparts ($M=.07$, $SD=.17$). This might be due either to the motivating effect of the interviewer or to the introduction the respondents received that encouraged them to spot errors. Further research is necessary to disentangle these effects.

However, error quantity does not necessarily indicate that one method is superior to the other (Willis et al. 1999). The possibility of false positives remains (Conrad and Blair 2009). The more important distinction for us lies in the error types a method reveals and whether these match across methods.

Table 2.1. Error Types Occurring in Item Battery on Specific National Pride

	CI (%)	OP (%)
Comprehension and Communication	77	79
Question Content		
Vague topic/unclear question (e.g., “Do you mean the question in a general sense? Or is it about a specific area?”)	12	1
Complex topic (e.g., “I can’t say that I’m proud of our history because we’ve got a shady past. There also was a history before the World Wars, which I’m quite proud of. It would misrepresent my answer if I would choose any answer value.”)	10	8
Topic carried over from earlier question (e.g., “This is similar to the question about economy.”)	3	1
Problematic term (e.g., “I didn’t initially know how to interpret ‘groups of society.’ This was unclear.”)	6	0
Term easily misunderstood (e.g., “Social security benefits? Do you mean education?”)	2	0
Inappropriate term (e.g., “The term pride bothers me a lot in this context.”)	8	5
Wrong underlying assumption (e.g., “I can’t be proud of something that I didn’t achieve myself.”)	8	27
Missing relevance for the respondent (e.g., “I have no interest in sports.”)	12	34
Action Code: Changing phrasing of question (e.g., “I’ll answer the question with how ‘content’ I am.”)	2	0
Question Structure		
Problematic phrasing of question (e.g., “I don’t understand the formulation of the question.”)	2	1
Several questions: Differing knowledge (e.g., “I can’t remember any artist. Regarding literature, the brothers Grimm cross my mind, and currently Frank Schätzing.”)	6	1
Undefined Reference Period (e.g., “Which German history? The last 100 years, the last 200 years? Where does it start?”)	5	1
Retrieval	10	17
Information unavailable (e.g., “I don’t have a clue about arts and literature.”)	10	17
Judgment	4	2
Social desirability (e.g., “The statement ‘proud of being German’ sounds Nazi.”)	4	2
Response Selection	10	3
Missing response categories (e.g., “There is no middle-category. I need a middle-category.”)	3	0
Wrong category titles (e.g., “I would prefer ‘very good’ to ‘very proud’.”)	4	3
Action code: Switching of answer value (e.g., “I’ve reconsidered and chose ‘somewhat proud’ instead of ‘very proud’.”)	2	0

Notes: Percentages are based on the number of all identified errors in each method (error total: CI=95; OP=112) regarding the whole item battery. CI = cognitive interviewing; OP = online probing.

Which error types are identified? As Table 2.1 shows, the majority of errors found in both methods belong to Tourangeau's comprehension and communication stage (CI: 77 percent, OP: 79 percent). This is in line with previous comparisons between CI and other pretesting methods (e.g., Rothgeb et al. 2007). However, both methods differ in the frequency of errors related to the retrieval stage (CI: 10 percent, OP: 17 percent). The increased error prevalence related to the retrieval stage during OP clearly flags the items "arts and literature" (37 percent of OP retrieval-related incidences) and "history" (26 percent of OP retrieval-related incidences) as problematic. In the anonymity of the web, the web respondents seemingly dare to admit more often their ignorance of a topic.

In total, CI uncovered 17 error types and OP found 12 error types. In general, both methods overlap in most error types that were detected. For example, both methods agree that the item battery is problematic because several respondents perceived the topic of some items such as "social security system" or "economic achievements" as too complex to generate an answer (code "complex topic").

Additionally, the respondents of both methods objected to the word "pride" in the question wording. Either they disliked the wording (code "inappropriate term"), they saw it as socially undesirable to be proud of their country (code "social desirability"), or they criticized the underlying assumption of the question (code "underlying assumption"). Such problems were uncovered in both methods, for example, for the items "political influence in the world" and "history."

However, the relative importance of some error types differed across methods. Cognitive interviewees suggested more often than OP respondents that an item was too vague and that they would need further information to answer it (code "vague topic," CI: 12 percent, OP: 1 percent). This discrepancy might be due to false positives in CI. Interviewer presence or encouragement to spot errors might tempt CI participants to voice problems where OP respondents still seem to make sense of the question.

Due to the larger sample size, OP showed a prevalence of some error codes. Around one-third of all OP error incidences referred to the “missing relevance” of a subject for the respondents. Thus, the items “sport” (16 times) and “arts and literature” (13 times) were particularly flagged as problematic.

Despite the smaller sample size, CI found error types that OP missed. Only CI respondents mentioned the error codes “problematic term,” “term easily misunderstood,” and “missing response category,” and the two action codes “switching of answer value” and “changing phrasing of the question.” It is possible that some of these error types might be “false positives.” The small CI sample size prevents a clear distinction between “false positives” and real errors.

So far, the aggregated results for the entire item battery have been presented. A further interesting question is the overlap in error detection between both methods at the item level. Both methods agreed that each item is somewhat problematic because each method found at least one error type at each item. However, there were discrepancies in the number of error types found: CI spotted additional error types at nine items and OP revealed error types that were not mentioned during CI at seven items. The most extreme example is the item “fair and equal,” where OP revealed two but CI found eight different error types. The online section contains a table (Table S2.1) with the results at item level. The question remains whether both methods suffer from an insufficient sample size that could create such a mismatch. This means: With increased sample sizes in both methods, the mismatch would possibly decrease and false positives easier identified.

How does the divergent level of interactivity affect the performance of both methods on finding errors? Another feature of CI was the possibility of interviewers to react with emergent probes in case of insufficient answers. Thus, additional errors could be spotted that had not been uncovered through standardized probes (8 percent of CI error codings). For

example, emergent probing showed that respondents struggled with the term “groups in society” at the item “fair and equal.”

CI: Where is the problem?

P: What is being asked anyway? Groups, do you mean in Germany or Europe? What are groups in society? Or are these my neighbors, e.g., Turks, Italians? This is why I do not know. (P4, CI)

In OP, a nonresponse or insufficient answer triggered a repetition of the probe with an additional motivational sentence. These conditional probes had to be programmed beforehand and were thus not tailor-made for a particular problem. In total, 5 percent of OP error incidences were prompted through a conditional probe.

Remarkably, the spontaneous remarks of CI participants turned out to be more important for the uncovering of errors than emergent probing. In total, 22 percent of CI error codings came from spontaneous comments about the items before the standardized probe was administered. We conclude that the increased level of interactivity improved the CI performance with regard to error detection.

2.4.3 Do The Methods Uncover Similar Themes? The Theme Perspective

In a second step, this study compared a subset of three items from a theme perspective. The theme perspective does not look for overt errors but for more implicit differences in the themes respondents think of when they answer a question. Based on the probe answers from CI and OP, a separate coding schema for each item was developed. The analyzed items were:

- a. “pride in the social security system”
- b. “pride in the achievements in arts and literature”
- c. “pride in the fair and equal treatment of all groups in society”

We selected these items because Latcheva’s study had identified them as problematic due to multiple respondent perspectives (2011). The coding schemas captured the answers to these probes:

“What particular **social security benefits** did you have in mind when you were answering the question?”

“What particular **achievements in the arts and literature** did you have in mind when you were answering the question?”

“What particular **groups in society** did you have in mind when you were answering the question?”

The same coding procedure as in the error perspective applied. The intercoder agreement varied between 91 percent and 100 percent.

How many themes are identified? We compare all respondents that gave a substantive probe answer (CI: for all items $N=20$, OP: social security” $N=95$, “arts and literature” $N=92$, “fair and equal” $N=95$). Contrary to the error perspective, OP respondents were nearly as productive as CI respondents because they stated 1.5–1.8 themes on average whereas CI respondents mentioned 1.6–1.9 themes.

Do CI and OP methods uncover similar themes? We are particularly interested in whether both methods identify the same themes because divergent themes would indicate interpretation differences between the respondents of both methods. As respondents could mention several themes, multiple coding applied. Any theme that was not mentioned by at least 5 percent of respondents in at least one method was summarized in the “others” category.

When respondents were asked what benefits they were thinking of when they rated their pride in Germany’s social security system (see Table 2.2), the CI and OP participants were mostly thinking of unemployment, health care, welfare, retirement, and family benefits. The code “no particular benefits” was the only code not mentioned during CI but provided by 6 percent of OP respondents. Answering “no particular benefit” might indicate satisficing (Krosnick 1991). It is harder to think of a particular benefit than to answer that one was

thinking of nothing in particular, even though the latter might also be true. Nonetheless, we conclude that both methods had largely overlapping results for the item “social security.”

Table 2.2. Benefits Mentioned for the Item “Social Security”

Social Security Benefits	CI (<i>N</i> = 20) (%)	OP (<i>N</i> = 95) (%)
Welfare	35	21
Unemployment	60	53
Health care	35	34
Long-term care	5	2
Retirement	25	21
Family	20	25
Educational	5	10
No particular	0	6
Others	0	9

Note: CI = cognitive interviewing; OP = online probing.

A similar situation applied to the item “arts and literature” (see Table 2.3). Respondents of CI and OP were thinking of similar achievements when they rated their pride in Germany’s *achievements in arts and literature*. Most respondents in both methods referred to achievements in literature and the visual arts. Some respondents in CI and OP thought of achievements in music or performing arts. Once again, only OP respondents thought of “no specific achievements” and might have satisfied.

Table 2.3. Achievements Mentioned for the Item “Arts and Literature”

Achievements	CI (<i>N</i> = 20) (%)	OP (<i>N</i> = 92) (%)
Literature	85	62
Music	15	14
Performing arts	5	5
Visual arts	50	50
No specific achievements	0	14
Others	0	5

Note: CI = cognitive interviewing; OP = online probing.

The situation differs for the item “fair and equal” (see Table 2.4). Most respondents of both methods mentioned “ethnic minorities,” people discriminated against because of their “sexual orientation,” or people discriminated against because of their “financial situation.” However, the relative importance of the themes differs. Far more CI respondents were thinking of ethnic minorities than OP participants (CI: 70 percent, OP: 28 percent). The local

recruitment of CI participants in the urban area of Mannheim could explain this difference. CI respondents could think more often of ethnic minorities because Mannheim is a multicultural city with 39 percent of inhabitants with a migration background (Kommunale Statistikstelle Stadt Mannheim 2014). As a CI respondent remarks:

P: As a resident of Mannheim, I directly have to think of our Bulgarians who are wrongfully accused, made responsible, for everything that is going wrong. This is what I was spontaneously thinking of. (P3, CI)

OP respondents are more geographically dispersed and might, therefore, give a more realistic presentation of which groups of society Germans in general think of when they answer this item.

Additionally, OP respondents mentioned several groups that were not revealed during CI. For example, 10 percent of OP respondents thought of the discrimination of women (theme “gender”), and 12 percent of OP participants referred to age-related discrimination (theme “age”). Due to the larger sample size, OP uncovered more themes for this item than CI. This might be an example of the limitation of CI sample sizes.

Table 2.4. Social Groups Mentioned for the Item “Fair and Equal”

Social Groups	CI (N = 20) (%)	OP (N = 95) (%)
Ethnic minorities	70	28
Sexual orientation	25	15
Financial situation	25	41
Religion	5	13
Family	5	1
Gender	0	10
Health	10	12
Age	0	12
Minorities in general	5	1
All	0	7
None	5	8
Others	0	7

Note: CI = cognitive interviewing; OP = online probing.

How does the divergent level of interactivity affect the performance of both methods in finding themes? The influence of interactivity was not as clear cut between both methods in the theme perspective as in the error perspective. For the item “social security,” emergent probing did not reveal any new information during CI, and only 2 percent of OP codings were

due to conditional probing. For the item “arts and literature,” emergent probing resulted in 7 percent of CI codings; the corresponding OP figure is 6 percent. Finally, for the item “fair and equal,” emergent probing contributed with 7 percent to CI codings; for OP, the corresponding figure is 2 percent. In contrast to the error perspective, the role of spontaneous remarks from CI respondents was negligible (one incidence). In sum, no clear difference in the influence of interactivity can be detected between both methods in the theme perspective.

2.5 DISCUSSION

2.5.1 Evaluation of Research Questions

First, we found indications that the probe response quality was higher in CI because CI respondents had lower nonresponse and longer responses. In contrast, the nationwide sampling in OP prevented any local bias in results.

Second, the 20 CI participants uncovered numerous error types and themes and were slightly more productive in both perspectives than OP respondents. However, the lower OP response quality was compensated through the larger sample size. OP participants revealed various error types and themes. The large sample size also allowed judging the prevalence of some error types, which is an important tool to avoid false positives. Although the methods differed in several aspects, they had an extensive overlap of results in both perspectives.

Third, the impact of interactivity on the performance of both methods differed across perspectives. On the one hand, the high interactivity level of CI improved its performance in error detection. This was largely due to spontaneous remarks of respondents. So far, OP cannot record spontaneous remarks, and the conditional probes could not compensate for the absence of an interviewer. On the other hand, the divergent level of interactivity had no clear impact on the performance of both methods during the theme perspective. The necessary or allowed degree of interactivity differs with the research goal. Interactivity allows for intensive

interviewing, but a high level of non-standardized probing could also threaten the comparability of results (Conrad and Blair 2009).

2.5.2 Researcher Burden in Both Methods

Researcher need to consider several practical factors when choosing between methods. The researcher burden depends on the level of experience of the team and its skill sets. For CI, the participants need to be recruited and the interviewers might still need training. As CI often tests many items, interviews take between 60 and 90 minutes. After data collection, the interviews may need to be transcribed.

In contrast, for OP it is necessary to search for a panel provider and to program the web survey. The latter can be outsourced, but it will increase costs. Although the responses do not need to be transcribed, data analysis will take some time due to the increased sample size. Future studies should thus look into the potential of (semi-)automated coding of probe answers. Further information on differences between the methods can be found in Table S2.2 of the online appendix

2.5.3 Limitation

Our research goal was to compare the “usual” approaches of CI and OP. Two limitations follow out of this research goal: First, OP respondents did not receive the same briefing as CI participants, who were encouraged to spot errors. This briefing, together with the presence of an interviewer, is likely to explain the higher productivity of CI respondents in finding errors. Second, the divergent sample sizes might be problematic because we compared 20 cognitive interviews with 532 online respondents that were split into five groups. A larger CI sample size would probably increase the variety of detected themes and errors. Given these limitations, future research should test framing web probing in the context of error detection and also choosing larger/equal sample sizes.

2.5.4 Optimal Application

The results of this study show that each method has its own strengths but also that the methods can effectively complement each other. Thus, CI and OP could be combined in a single study.

Thanks to its interactivity, CI can be particularly useful at the exploratory stage of a research project. CI allows one to gain a deep understanding of respondents' thoughts and helps assess whether the developed items measure what they are supposed to measure. Given the motivational impact of the interviewer, it is possible to probe a high number of items. Therefore, CI especially lends itself as a pretesting tool for newly developed questionnaires. The CI pretest could exclusively aim at error detection or it could also focus on respondents' interpretation of items in general (Miller et al. 2014).

OP could be implemented at any stage of the survey process. As a pretesting tool, it could become particularly useful once the researcher has a thorough understanding of the studied phenomena. Typically, only a subset of potentially problematic items would be probed though. The large OP sample sizes allow for assessing the prevalence of error types or themes mentioned by respondents and for studying certain subgroups of respondents and answer combinations—aspects that might be difficult in CI. During data collection, OP might serve the purpose of quality control, especially if the actual study itself is a web survey. After data collection, OP can become a powerful tool for guiding analysis and interpretation of surprising quantitative results. However, to optimally use OP, some knowledge of the subject matter is necessary to make an informed decision as to which probe type to apply. CI can guide this process and thus provide the basis for OP studies in larger and more heterogeneous (e.g., nationwide) respondent groups.

References

- Beatty, Paul C. and Gordon B. Willis. 2007. "Research Synthesis: The Practice of Cognitive Interviewing." *Public Opinion Quarterly* 71:287–311.
- Behr, Dorothée, Michael Braun, Lars Kaczmirek, and Wolfgang Bandilla. 2014. "Item Comparability in Cross-national Surveys: Results from Asking Probing Questions in Cross-national Web Surveys about Attitudes towards Civil Disobedience." *Quality & Quantity* 48:127–48.
- Blair, Jonny and Frederick G. Conrad. 2011. "Sample Size for Cognitive Interview Pretesting." *Public Opinion Quarterly* 75:636–58.
- Braun, Michael, Dorothée Behr, Lars Kaczmirek, and Wolfgang Bandilla. 2015. "Evaluating Cross-national Item Equivalence with Probing Questions in Web Surveys." Pp. 184–200 in *Improving Survey Methods: Lessons from Recent Research*, edited by U. Engel, B. Jann, P. Lynn, A. Scherpenzeel, and P. Sturgis. New York: Routledge, European Association of Methodology.
- Conrad, Frederick G. and Johnny Blair. 2009. "Sources of Error in Cognitive Interviews." *Public Opinion Quarterly* 73:32–55.
- DeMaio, Theresa J. and Ashley Landreth. 2004. "Do Different Cognitive Interview Techniques Produce Different Results?" Pp. 89–108 in *Methods for Testing and Evaluating Survey Questionnaires*, edited by S. Presser, J. M. Rothgeb, M. P. Couper, J. T. Lessler. Hoboken, NJ: John Wiley & Sons.
- Fleiß, Jürgen, Franz Höllinger, and Helmut Kuzmics. 2009. "Nationalstolz zwischen Patriotismus und Nationalismus?" *Berliner Journal für Soziologie* 19:409–34.

- Kommunale Statistikstelle Stadt Mannheim. 2014. "Einwohner mit Migrationshintergrund in kleinräumiger Gliederung." Statistische Daten 3/2014. Mannheim, Germany: Kommunale Statistikstelle Stadt Mannheim. Retrieved September 18, 2014 (https://www.mannheim.de/sites/default/files/page/2188/d201403_migrationshintergrund_2013.pdf).
- Krosnick, John A. 1991. "Response Strategies for Coping with the Cognitive Demands of Attitude Measures in Surveys." *Applied Cognitive Psychology* 5:213–36.
- Latcheva, Rossalina. 2011. "Cognitive Interviewing and Factor-Analytic Techniques: A Mixed Method Approach to Validity of Survey Items Measuring National Identity." *Quality & Quantity* 45:1175–99.
- Miller, Kristen, Valerie Chepp, Stephanie Willson, and Jose L. Padilla. 2014. *Cognitive Interviewing Methodology*. New York: John Wiley & Sons.
- Murphy, Joe, Michael Keating, and Jennifer Edgar. 2013. "Crowdsourcing in the Cognitive Interviewing Process." Paper presented at the FCSM Research Conference, Washington, DC, November 4–6.
- Prüfer, Peter and Margrit Rexroth. 2005. "Kognitive Interviews." *ZUMA How-to-Reihe 15*. Retrieved September 18, 2014 (http://www.gesis.org/fileadmin/upload/forschung/publikationen/gesis_reihen/howto/How_to15PP_MR.pdf).
- Rothgeb, Jennifer, Gordon Willis, and Barbara Forsyth. 2007. "Questionnaire Pretesting Methods: Do Different Techniques and Different Organizations Produce Similar Results?" *Bulletin de Méthodologie Sociologique* 96:5–31.
- Smith, Tom W. and Lars Jarkko. 2001. *National Pride in Cross-national Perspective*. Chicago, IL: National Opinion Research Center. Retrieved September 20, 2014 (www.issp.org/documents/natpride.doc).

- Tourangeau, Roger, Lance J. Rips, and Kenneth Rasinski. 2000. *The Psychology of Survey Response*. Cambridge: Cambridge University Press.
- Willis, Gordon B. 2005. *Cognitive Interviewing: A Tool for Improving Questionnaire Design*. Thousand Oaks, CA: Sage.
- Willis, Gordon B., Susan Schechter, and Karen Whitaker. 1999. "A Comparison of Cognitive Interviewing, Expert Review, and Behavior Coding: What do They Tell Us." In *Proceedings of the Section on Survey Research Methods, American Statistical Association*, Alexandria, VA: American Statistical Association:28–37.

Appendix A. Error Types

Table S2.1 shows the overlap in error detection between CI and OP at the item level. Both methods agreed that each item is somewhat problematic because each method found at least one error type at each item. However, there were discrepancies in the number of error types found: CI spotted additional error types at nine items and OP revealed error types that were not mentioned during CI at seven items. The most extreme example is the item “fair and equal,” where OP revealed two, but CI found eight different error types.

Table S2.1. Variety of Occurring Error Types per Item and Method

Items from ISSP Item Battery on Specific National Pride	CI	OP
Democracy	5	7
Political influence	5	4
Economic achievements	3	5
Social security system	5	2
Scientific and technological achievements	3	3
Achievements in sports	4	4
Achievements in the arts and literature	7	6
Armed forces	6	3
History	5	3
Fair and equal treatment	8	2

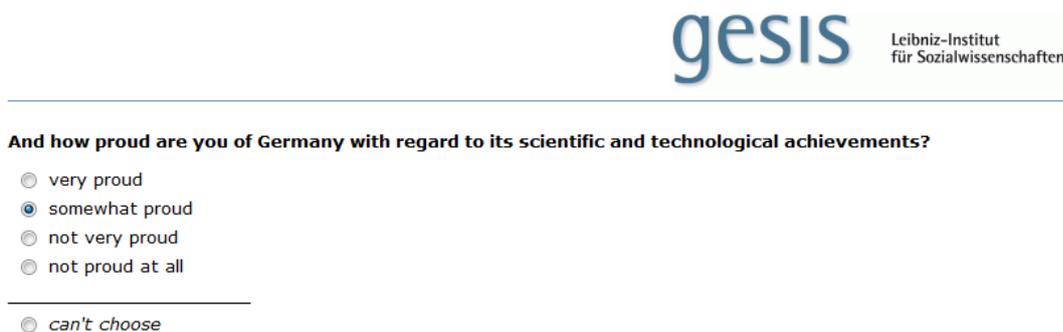
Note: CI = cognitive interviewing; OP = online probing.

Appendix B. Screenshots of Probes

EXAMPLE 1: ITEM “SCIENTIFIC AND TECHNOLOGICAL ACHIEVEMENTS”

After the first item probed in OP, respondents received a category-selection probe.

(Please note: The items and probes in the original survey were in German.)



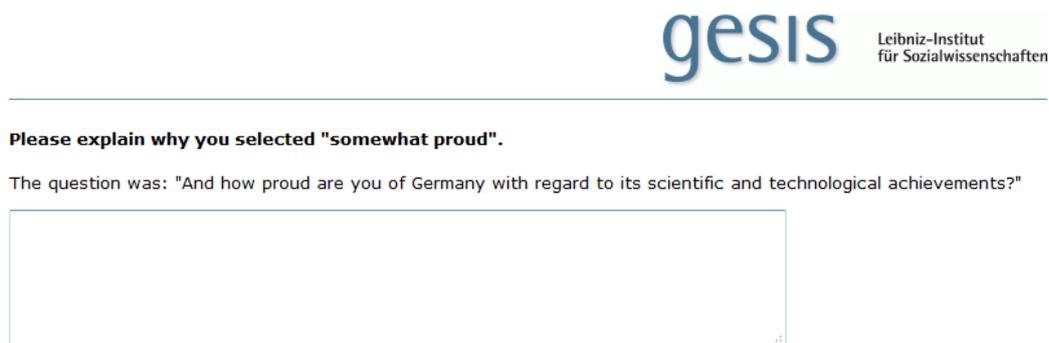
gesis Leibniz-Institut
für Sozialwissenschaften

And how proud are you of Germany with regard to its scientific and technological achievements?

very proud
 somewhat proud
 not very proud
 not proud at all

can't choose

Figure S2.1. Screenshot of Closed Item “Scientific and Technological Achievements”



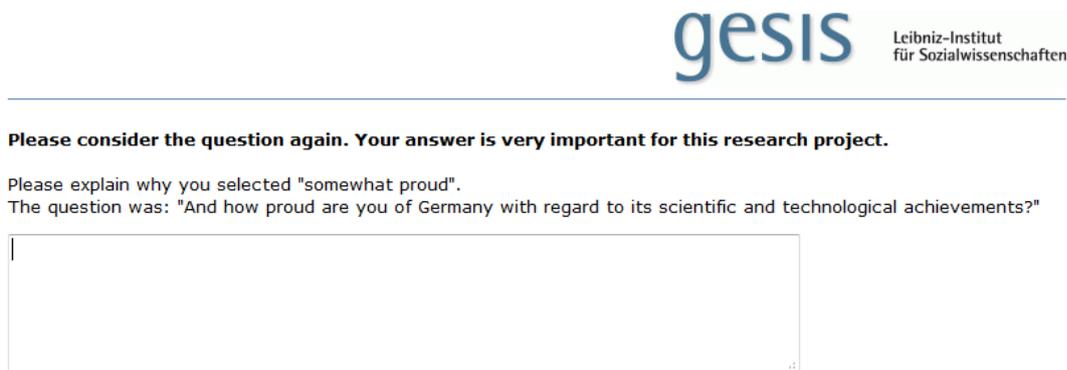
gesis Leibniz-Institut
für Sozialwissenschaften

Please explain why you selected "somewhat proud".

The question was: "And how proud are you of Germany with regard to its scientific and technological achievements?"

Figure S2.2. Screenshot of Category-selection Probe

In case of a nonresponse, the probe was repeated with an additional motivational sentence:



gesis Leibniz-Institut
für Sozialwissenschaften

Please consider the question again. Your answer is very important for this research project.

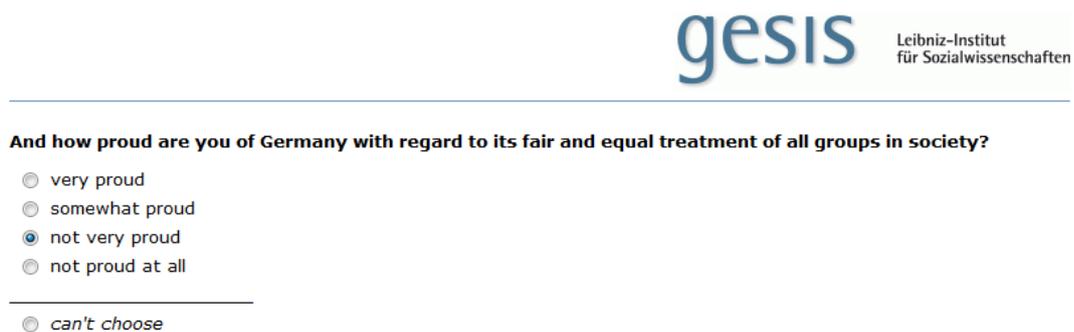
Please explain why you selected "somewhat proud".

The question was: "And how proud are you of Germany with regard to its scientific and technological achievements?"

Figure S2.3. Follow-up Probe in Case of a Nonresponse

EXAMPLE 2: ITEM “FAIR AND EQUAL”

For the second item probed in OP, the respondents received a category-selection probe and also a specific probe. In case of probe nonresponse, a nonresponse follow-up (conditional probing) would also appear after the category-selection probe and after the specific probe (not depicted here).



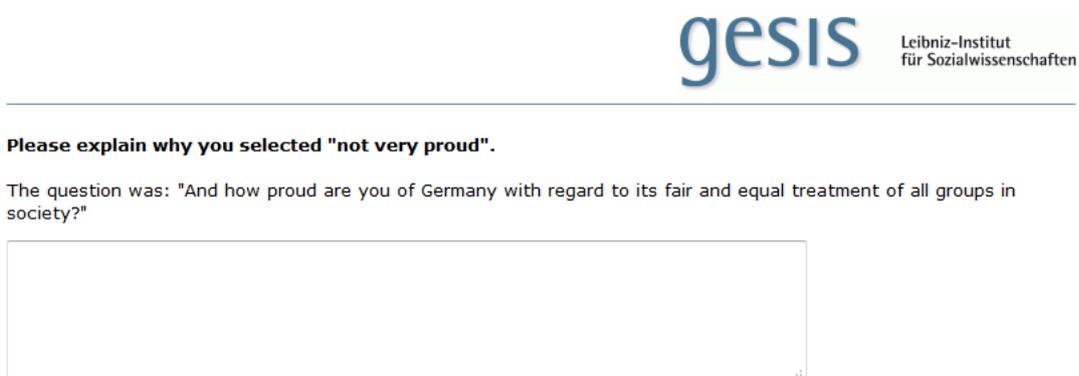
gesis Leibniz-Institut
für Sozialwissenschaften

And how proud are you of Germany with regard to its fair and equal treatment of all groups in society?

- very proud
- somewhat proud
- not very proud
- not proud at all

- can't choose*

Figure S2.4. Screenshot of Closed Item “Fair and Equal”



gesis Leibniz-Institut
für Sozialwissenschaften

Please explain why you selected "not very proud".

The question was: "And how proud are you of Germany with regard to its fair and equal treatment of all groups in society?"

Figure S2.5. Screenshot of Category-selection Probe



gesis Leibniz-Institut
für Sozialwissenschaften

What particular groups in society did you have in mind?

The question was: "And how proud are you of Germany with regard to its fair and equal treatment of all groups in society?"

Figure S2.6. Screenshot of Additional Specific Probe

Appendix C. Key Characteristics of CI and OP Studies

Table S2.2 Key Characteristics of CI and OP Studies

Key Characteristic	CI	OP
Number of respondents	20	532
Duration of survey/interview	Around 1 hour	Around 15 minutes
Length of questionnaire	45 items in total 27 items probed per respondent	26 items in total 8 items probed per respondent
Timing: Field period	9 days for 20 interviews (22.–30.04.2013)	8 days for 532 respondents (16.–23.09.2013)
Timing: Transcription of results	60 hours (3 hours for each interview)	Not necessary
Representation	Local sample	National sample
Number of questions that were probed so far (Item/probe numbers based on experience)	Potentially higher Usually probing of around 30 items with anticipated and emergent probes	Potentially lower So far, between 8 and 9 probes per respondent per 15-minute survey
Costs	2,140 € for 20 respondents	1,999 € for 532 respondents
	Includes: <ul style="list-style-type: none"> • Incentive of 30 € per participant including traveling expenses • Transcription of results • Basic error analysis • Report 	Includes: <ul style="list-style-type: none"> • Incentives of 1.50 € per respondent Not included: <ul style="list-style-type: none"> • Programming of web survey • Analysis • Report
Costs per respondent	107 €	3.76 €

Note: CI = cognitive interviewing; OP = online probing.

3. Necessary but Insufficient: Why Measurement Invariance Tests Need Online Probing as a Complementary Tool⁴

Abstract

Constructive patriotism and nationalism are two important concepts in the study of national identity. The popularity of both concepts in cross-national studies also stems from the work of Davidov (2009, 2011), who developed metric invariant measures of constructive patriotism and nationalism. This allows for a comparison of structural relationships, such as regression coefficients, but prohibits a cross-national comparison of latent means. The arrival of the 2013 ISSP Module on National Identity has given rise to a reassessment of both constructs and a push to understand why scalar invariance cannot have been achieved. Using the example of constructive patriotism and nationalism, this article shows how the combination of measurement invariance tests with multi-group confirmatory factor analysis (MGCFA) and online probing (OP) can uncover and explain issues related to cross-national comparability.

⁴ Preliminary results were presented at the XVIII ISA World Congress of Sociology, July 13-19, 2014, Yokohama, Japan, and at the ESA RN21 / EQMC Conference, October 24-25, 2014, Mannheim, Germany.

3.1 INTRODUCTION

With the proliferation of cross-national surveys, such as the International Social Survey Program (ISSP) or the European Social Survey (ESS), access to cross-national data sets has been tremendously facilitated. One precondition for the analysis of such data is the assessment of their cross-national comparability. Two research traditions can be distinguished in this context: quantitative and qualitative traditions. In the quantitative tradition, comparability is often assessed with measurement invariance tests that use multi-group confirmatory factor analysis (MGCFA) (Jöreskog 1971). This approach can test the cross-national comparability of numerous countries. The testing strategy is straightforward and is implemented using analysis software, such as Mplus, making it a handy control instrument for researchers who are interested in analyzing secondary data. However, if these tests fail to establish cross-national comparability, this approach struggles to explain the existing noninvariance. By contrast, researchers who use the qualitative approach will most likely conduct cognitive interviews (CIs) (Miller et al. 2011) or online probing (OP) (Braun et al. 2014). These methods primarily seek to uncover the causes for the lack of comparability of items, and they often reveal unexpected reasons. The drawbacks of these methods are the necessity of collecting data and the work-intensive analysis (Meitinger and Behr, forthcoming), which limit the analysis to a small set of countries. Much can be learned through a combined approach of both perspectives.

This article closes a research gap by simultaneously applying MGCFA and OP to assess the comparability of constructive patriotism and nationalism constructs. The article will introduce MGCFA and OP, followed by a short discussion of constructive patriotism and nationalism, which will include a review of previous research. The constructs' comparability will first be assessed with MGCFA and then with OP. Previous research has indicated that the items that measure constructive patriotism are especially error prone (Latcheva 2011); as such, we will focus on these items with regard to OP. Finally, the conclusions from both

methods will be compared, and optimal research strategies that combine the two methods' respective strengths will be presented.

3.1.1 The Quantitative Approach: Tests of Measurement Invariance

When working with quantitative cross-national data, ensuring that the measurement is invariant across countries is necessary (Hui and Triandis 1985). Otherwise, cross-national studies run the risk of misinterpreting ambiguous and erroneous data as “real” substantive differences across countries (Steenkamp and Baumgartner 1998).

Various techniques have been developed to test measurement invariance (Davidov et al. 2014). We use MGCFA for two reasons. First, MGCFA (Jöreskog 1971) is one of the most powerful approaches for measurement invariance tests (Meuleman 2012). Second, we build on Davidov's work (2009; 2011) on the measurement invariance of constructive patriotism and nationalism. To increase our results' comparability, we choose the same methodological approach.

Most researchers conduct three tests of comparability when applying MGCFA: *configural*, *metric*, and *scalar* invariance tests (Braun and Johnson 2010; Vandenberg and Lance 2000). The three tests are nested with configural invariance providing a test for the lowest level and scalar invariance providing a test for the highest level of invariance.

Configural invariance concerns whether all countries have the same factor structure, that is, if all items have the same configuration of salient and non-salient factor loadings in all countries (Horn and McArdle 1992). If configural invariance is established, the latent concept can be meaningfully discussed in all countries (Davidov et al. 2014). However, the respondents can still answer items differently because factor loadings may vary (Steenkamp and Baumgartner 1998).

The test for metric invariance addresses this issue by requiring equal factor loadings across countries (Rock, Werts, and Flaughner 1978). If metric invariance is supported, exploring cross-national structural relationships, such as regression coefficients, with other constructs is possible (Steenkamp and Baumgartner 1998). Since the requirements of equal factor loadings for all items might be challenging in cross-national comparisons, several researchers have advocated for *partial metric invariance*. Cross-national comparisons are acceptable if all constructs are measured with at least two items with equal factor loadings (Byrne, Shavelson, and Muthén 1989; Steenkamp and Baumgartner 1998).

However, many cross-national researchers aim to compare mean values across countries. As a systematic bias might affect the mean values (Meredith 1993), testing for scalar invariance is necessary. Scalar invariance tests additionally require equal intercepts. If full scalar invariance does not apply, opting for *partial scalar invariance* is another possibility (Steenkamp and Baumgartner 1998).

Usually the different levels of invariance tests are conducted using a bottom-up approach, with the acceptance of lower-level invariance tests as a precondition for conducting higher-level tests. Goodness-of-fit (GOF) indices, such as the *root mean square error of approximation* (RMSEA, Browne and Cudeck 1992) and the *comparative fit index* (CFI, Bentler 1990), are used to assess the model fit of the baseline (configural) model. Hu and Bentler (1999) suggest CFI values of at least .95 and RMSEA values below .06 for a good model fit, and RMSEA values below .08 for an acceptable model fit. If configural invariance is achieved, the baseline model can be compared with more restricted models. Previous studies have often used the chi-square difference test for this purpose. As this test is sensitive to large sample sizes (Cheung and Rensvold 2002; Davidov 2011), Chen (2007) instead proposed using the difference in the CFI and RMSEA values of the different test levels, ΔCFI and ΔRMSEA , to assess model fit. A change of more than .01 for CFI and .015 for RMSEA indicates problematic values (Cieciuch and Davidov, forthcoming).

If the GOF values are unsatisfactory, MGCFA provides *modification indices* (MIs) (Steenkamp and Baumgartner 1998), which help the researcher decide which parameters to free to improve the model fit. As the MIs are also sensitive to sample size (Cheung and Rensvold 2002), Saris and colleagues (2009) suggested considering the following three criteria to improve model fit: the MI for a parameter, the power of the MI test, and the *expected parameter change* (EPC), which estimates the degree of the parameter's misspecification. If these values are large, the researcher should consider model respecification. With *Jrule* software for Mplus, these values can be easily obtained (Oberski 2014).

Although MGCFA provides the researcher with indications of troublesome items, it does not explain *why* certain items are problematic in cross-national comparisons. Furthermore, the provision of MIs and EPCs might tempt researchers to provide substantive ad hoc explanations for noninvariance. However, when using this approach, determining whether measurement invariance is missing because of a methodological artifact or because of different realities is impossible.

Several quantitative approaches exist that aim to reveal the sources of noninvariance. For example, on the micro level, the multiple indicators multiple causes (MIMIC) model tests whether the item is affected by individual variables, such as age or gender, and controls for this differential item functioning (Davidov et al. 2014). On the macro level, multilevel structural equation models (MLSEMs) try to explain noninvariance by introducing conceptual predictor variables in a multilevel analysis (Davidov et al. 2012). However, for an accurate estimation, the sample should consist of at least 50 countries (Meuleman and Billiet 2009); therefore, such estimation is not available for studies that compare a small set of countries. Additionally, the MIMIC and MLSEM approaches need a priori hypotheses about cultural differences or the reasons for bias (van de Vijver 2011). The study will only be as good as the researchers' capabilities to discover the correct explanations for noninvariance to include in

the analysis and, of course, the availability of the corresponding data. However, previously unknown and surprising causes might exist. Finding these unexpected causes is one of the major advantages of qualitative approaches.

3.1.2 The Qualitative Approach: Online Probing in a Cross-national Context

Different qualitative approaches can evaluate the cross-national comparability of items. During traditional CIs, survey questions are administered to respondents “while collecting additional verbal information about survey responses, which is used to evaluate the quality of the response or to help determine whether the question is generating the information that its author intends” (Beatty and Willis 2007: 287). Interviewers ask follow-up questions called “probes” to retrieve additional information. For example, *category-selection probes* ask why a certain answer category was chosen, and *specific probes* require additional information on a particular detail in the question (Prüfer and Rexroth 2005; Willis 2005). Probing is also a powerful tool for detecting instances of *silent misinterpretation*, where respondents are unaware that they have misunderstood the item (DeMaio and Rothgeb 1996).

A rather recent approach is OP, which applies probing techniques from CIs in web surveys. OP combines qualitative insights from CIs with large sample sizes in several countries. As all respondents receive the same probe, the procedure is highly standardized (Braun et al. 2014). It thus avoids harmonization issues, a challenge in cross-national CIs due to, e.g., the varying levels of interviewers’ skills (Gray and Blake 2015). The large sample size increases the generalizability of results, allows for an evaluation of the prevalence of problems or themes, and can explain the response patterns of specific subpopulations (Braun et al. 2014).

3.1.3 Constructive Patriotism and Nationalism: Theory and Empirical Approaches

Although national identity constitutes a relevant field of research for the social sciences (Latcheva 2011), its definition and conceptualization remain controversial (Davidov 2009). Several studies distinguish between two elements of national identity. Adorno and colleagues (1950) first introduced the idea of differentiating between a love of one's country (genuine patriotism) and an uncritical attachment to one's country combined with a rejection of other nations (pseudo-patriotism). Kosterman and Feshbach (1989) also underlined that national identity should be seen as a multidimensional construct. The distinction was further developed by Schatz, Staub, and Lavine (1999; blind versus constructive patriotism) and Blank, Schmidt, and Westle (2001; nationalism versus constructive patriotism). Most prominently, in the German context, Blank and Schmidt (2003) juxtapose nationalism and constructive patriotism and see both as specific components of national identity. According to their perspective, nationalists idealize their nation, show feelings of national superiority, and have an uncritical acceptance of national authorities. Nationalists also suppress ambivalent attitudes toward the nation, tend to define their group based on descent, race or culture, and denigrate groups that they do not consider part of the nation. By contrast, constructive patriots reject an idealization of the nation. Their support for the nation depends on its alignment with humanistic and democratic principles. They value an advanced social system, are open to criticism and reject an uncritical acceptance of state authorities (Blank and Schmidt 2003; Davidov 2009).

3.1.4 Evaluation of Constructive Patriotism and Nationalism

A quantitative evaluation of the measurement of constructive patriotism and nationalism. Davidov (2009) adapted Blank and Schmidt's measure to the cross-national context using five items from the 2003 ISSP Module on National Identity. Nationalism was measured with two items (on a five-point scale ranging from "strongly disagree" to "strongly

agree”) that aim to capture feelings of national superiority. The three items evaluating one’s pride in the country’s democracy, its social security system, and the fair and equal treatment of all groups in society served as indicators of constructive patriotism (on a four-point scale ranging from “very proud” to “not proud at all”) (see Table 3.1). As both constructs are only measured with five items, measurement invariance tests must simultaneously assess constructive patriotism and nationalism. If tested separately, the metric invariance test will become unfeasible, as the models are either just identified (constructive patriotism) or not identified (nationalism).

Table 3.1. Items Measuring Nationalism and Constructive Patriotism in ISSP 2013

Factor	Item Name	Question Wording
NAT	V19: “More like us”	The world would be a better place if people from other countries were more like the [COUNTRY NATIONALITY].
	V20: “Better country”	Generally speaking, [COUNTRY] is a better country than most other countries.
COP	V25: “Democracy”	How proud are you of [COUNTRY] in the way democracy works?
	V28: “Social security”	How proud are you of [COUNTRY] in its social security system?
	V34: “Fair and equal”	How proud are you of [COUNTRY] in its fair and equal treatment of all groups in society?

Davidov (2009) used 34 countries to test the constructs for measurement invariance. He could establish metric invariance of both constructs, which allowed for a comparison of the constructs’ correlates but not their means. Additionally, Davidov (2011) investigated whether both ISSP measures are invariant over time, thus using data from 1995 ISSP and 2003 ISSP. He could confirm partial scalar invariance for 21 of the 22 countries, thus supporting a comparison of the constructs’ correlates and means over time.

A qualitative evaluation of the measurement of constructive patriotism. Previous qualitative research has indicated that several of these items are error prone. So far, two CI studies in Austria (Fleiß et al. 2009; Latcheva 2011) and one CI and OP study in Germany (Meitinger and Behr, forthcoming) tested these items and uncovered several issues: (1) Respondents in both countries indicated problems with the word “pride,” which distorts many respondents’ answers (Latcheva 2011; Meitinger and Behr, forthcoming). (2) German

respondents also struggled with the high complexity of terms such as “social security system” (Meitinger and Behr, forthcoming). (3) Several key terms (e.g., “all groups of society” and “democracy”) were not specifically formulated enough and led respondents to adopt different perspectives when they answered these items. For example, the term “social security system” triggered references to various benefits, comparisons with other countries, and current government policies (Latcheva 2011). Additionally, respondents associated a general value, political disenchantment, and current government policies with the term “democracy” (Fleiß et al. 2009; Latcheva 2011). Although these studies could uncover several issues with the tested items, they did not adopt a cross-national comparative perspective.

3.2 RESEARCH OBJECTIVE

The overarching goal of this article is to demonstrate how MGCFA measurement invariance tests and OP can be combined to assess the cross-national comparability of survey items. For this purpose, we evaluate the cross-national comparability of the measurement instruments for nationalism and constructive patriotism using MGCFA. Then, the OP results for the three items that measure constructive patriotism are presented, as these items were shown to be the most error prone (Latcheva 2011).

3.3 METHODS AND DATA

3.3.1 Multi Group Confirmatory Factor Analysis

For the MGCFA measurement invariance tests, we used the data set from the 2013 ISSP Module on National Identity (ISSP Research Group 2015) and limited our analysis to the five countries in our web survey: Germany ($N=1,717$), Great Britain ($N=904$), the U.S. ($N=1,274$), Mexico ($N=1,062$), and Spain ($N=1,225$). Following Davidov’s approach, we measured constructive patriotism and nationalism with the items presented in Table 3.1.

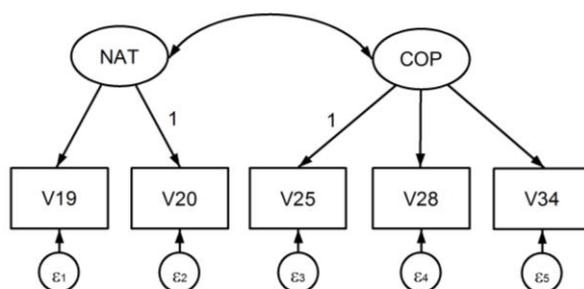


Figure 3.1. Confirmatory Factor Analysis of Nationalism (NAT) and Constructive Patriotism (COP)

3.3.2 Online Probing

The OP results were generated from a web survey conducted with 2,685 respondents in May 2014. The survey participants from Germany, Great Britain, Mexico, the U.S., and Spain were drawn from a non-probability online panel with quotas for age (18–30, 31–50, and 51–65), gender, and education (lower and higher). The survey replicated questions from the ISSP Module on National Identity. Since probing increases the response burden for respondents, the sample was randomly split in five groups of approximately 500 respondents each (approximately 100 respondents per country)⁵. All respondents answered each closed item on a separate screen. For each item, one-fifth of the respondents received an additional probe on a separate screen. In our OP study, we focused on the items that measured constructive patriotism because the previously discussed qualitative studies found that these three items were problematic (Latcheva 2011).

After the “democracy” item, a category-selection probe (Prüfer and Rexroth 2005) inquired why a certain answer category had been chosen. A specific probe that asked for additional information on a detail of the question followed the “social security” and “fair and

⁵ Unfortunately, this split condition prevented us from conducting the measurement invariance tests for constructive patriotism and nationalism with our web survey, since the nationalism items were answered by different respondents than the constructive patriotism items.

equal” items (Willis 2005). Figures 3.2-3.4 show screenshots of the three probes. Based on the probe answers, a separate coding schema was developed for all three items. A researcher coded all probe responses, and student assistants coded them a second time. Multiple coding was possible for all probes. The intercoder reliability was high (“democracy”: 94 percent; “social security”: 97 percent; “fair and equal”: 98 percent). Mismatched coding was discussed among the coding team.

Please explain why you selected "not very proud".

The question was: "How proud are you of America with regard to the way democracy works?"


 A screenshot of a survey question. The question text is "How proud are you of America with regard to the way democracy works?". Below the question is a large, empty rectangular text input field with a small cursor icon at the bottom right corner.

Figure 3.2. Category-selection Probe for the “Democracy” Item

What particular social security benefits did you have in mind when you were answering the question?

The question was: "And how proud are you of America with regard to its social security system?"


 A screenshot of a survey question. The question text is "And how proud are you of America with regard to its social security system?". Below the question is a rectangular text input field with a small cursor icon at the bottom right corner.

Figure 3.3. Specific Probe for the “Social Security” Item

What particular groups in society did you have in mind?

The question was: "And how proud are you of America with regard to its fair and equal treatment of all groups in society?"


 A screenshot of a survey question. The question text is "And how proud are you of America with regard to its fair and equal treatment of all groups in society?". Below the question is a rectangular text input field with a small cursor icon at the bottom right corner.

Figure 3.4. Specific Probe for the “Fair and Equal” Item

3.3.3 Comparison of ISSP and Web Survey Results

We consider the replication of the ISSP answer distribution of the constructive patriotism items in the web survey as a precondition for evaluating their cross-national comparability with our probes. Table 3.2 summarizes the means, standard deviations, and nonresponse rates for the 2013 ISSP and our web survey.

Table 3.2. Comparison of the Mean Values, Standard Deviations, and Nonresponse Rates between 2013 ISSP and Web Survey for the Items Measuring Constructive Patriotism

Country	2013 ISSP						Web Survey					
	Democracy		Social Security		Fair and Equal		Democracy		Social Security		Fair and Equal	
	Mean (<i>SD</i>)	NR (%)										
Germany	2.2 (.7)	7.3	2.0 (.7)	5.7	2.5 (.8)	13.2	2.4 (.8)	8.1	2.4 (.8)	6.5	2.7 (.8)	12.1
GB	2.1 (.7)	9.0	2.3 (.8)	7.2	2.1 (.8)	7.5	2.3 (.9)	6.8	2.5 (.9)	8.7	2.3 (.9)	11.8
Mexico	3.2 (.8)	1.2	3.1 (1.0)	2.3	3.0 (1.0)	2.9	3.3 (.7)	2.3	3.3 (.8)	1.9	3.1 (.8)	4.0
Spain	3.0 (.9)	2.5	2.3 (1.0)	.9	2.6 (1.0)	4.0	3.1 (.9)	1.2	2.1 (.9)	.5	2.9 (.9)	3.7
U.S.	2.0 (.8)	7.9	2.5 (.9)	7.1	2.3 (.9)	7.5	2.0 (.9)	4.3	2.4 (.9)	6.8	2.3 (.9)	7.5

Note: All items are measured on a four-point scale running from, 1 “very proud” to 4 “not proud at all.”

3.4 RESULTS OF MEASUREMENT INVARIANCE TESTS: MGCFA

In the first step, we tested for cross-national measurement invariance for constructive patriotism and nationalism using the 2013 ISSP data. We used the Mplus 7.31 software package. As the nonresponse rate was high in some countries (e.g., “fair and equal” item: 13.2 percent in Germany), we chose the full information maximum likelihood (FIML) estimator with raw data. The FIML is particularly well suited to account for missing data (Brown 2014)⁶.

3.4.1 Single-country Analysis

A preliminary step for measurement invariance tests involves establishing that the model fits well in each country (Byrne and van de Vijver 2010). We started our analysis by running a separate CFA in each country. The standardized factor loadings were sufficiently high in all

⁶ Continuous variables are a precondition for the FIML estimator. However, all data derived from Likert scale items are ordinal by definition. Davidov and colleagues (2011) were able to show that using Likert scales for measurement invariance tests that apply MGCFA is justifiable, which is also supported by a simulation study from De Beukelaer and Swinnen (2011). We re-estimated the model with the weighted least squares means- and variance-adjusted (WLSMV) approach, an estimator for ordinal data (Flora and Curran 2004). Both estimators arrive at similar conclusions. The results for the WLSMV estimator are included in the appendix (Tables S3.1, S3.2, and S3.2).

countries, reaching at least .50 (see Table 3.3). Additionally, the moderate correlation between the two latent factors supported a two-factor solution (see Table 3.4). The CFI values were above .95, which indicated a very good model fit in all countries. The RMSEA values suggested a very good model fit for Germany, Great Britain, and Spain but only an acceptable fit for the U.S. and Mexico. We accepted the single-country model for all countries without any modifications. However, given the elevated RMSEA values, we still inspected the MIs, the power of the MI test, and the EPC values for the U.S. and Mexico in Jrule. For the U.S., the parameter with the highest MIs and EPCs was a cross loading of the “fair and equal” item on the nationalism factor. Some U.S. respondents might perceive this item as representing patriotic values, and some might perceive that it reflects nationalistic attitudes. Perhaps some Americans feel superior to other countries because of their country’s focus on egalitarian issues. For Mexico, Jrule recommended freeing an error correlation between the “democracy” and “social security” items. Apparently, a cause other than the constructive patriotism factor explains part of the correlation between these two items.

Table 3.3. Single-country Analysis: Unstandardized and Standardized Factor Loadings and Standard Errors

<i>Country</i>	<i>N</i> → <i>V19</i>	<i>N</i> → <i>V20</i>	<i>CP</i> → <i>V25</i>	<i>CP</i> → <i>V28</i>	<i>CP</i> → <i>V34</i>
(a) Factor loadings on nationalism and constructive patriotism (unstandardized) (standard error in parentheses)					
1. Germany	.73 (.08)	1	1	.88 (.06)	.92 (.06)
2. Great Britain	.76 (.12)	1	1	1.04 (.10)	1.11 (.11)
3. Mexico	.81 (.07)	1	1	1.12 (.06)	.94 (.06)
4. Spain	.73 (.05)	1	1	.82 (.06)	.95 (.07)
5. U.S.	1.02 (.14)	1	1	1.10 (.11)	1.33 (.16)
(b) Factor loadings on nationalism and constructive patriotism (standardized) (standard error in parentheses)					
1. Germany	.65 (.04)	.91 (.05)	.69 (.02)	.63 (.02)	.60 (.03)
2. Great Britain	.61 (.05)	.87 (.07)	.64 (.04)	.58 (.04)	.64 (.04)
3. Mexico	.68 (.03)	.80 (.03)	.76 (.02)	.75 (.02)	.62 (.03)
4. Spain	.66 (.03)	.86 (.03)	.68 (.03)	.52 (.03)	.61 (.03)
5. U.S.	.62 (.05)	.66 (.05)	.52 (.04)	.50 (.04)	.61 (.04)

Note: N = nationalism factor; CP = constructive patriotism factor.

Table 3.4. Single-country Analyses: RMSEA, CFI, and Correlations between Nationalism and Constructive Patriotism and Error Correlations (Standard Errors in Parentheses)

<i>Country</i>	<i>RMSEA</i>	<i>CFI</i>	<i>Correlation</i>
1. Germany	.063 [.044; .084]	.982	N↔ CP: .36 (.03)
2. Great Britain	.000 [.000; .044]	1.000	N↔ CP: .38 (.05)
3. Mexico	.074 [.049; .101]	.982	N↔ CP: .54 (.04)
4. Spain	.055 [.032;.082]	.987	N↔ CP: .68 (.03)
5. U.S.	.075 [.052;.100]	.952	N↔ CP: .49 (.05)

3.4.2 Measurement Invariance Tests

We started by assessing the cross-national configural invariance. This baseline test evaluated whether all countries have the same factor structure. The GOF values confirmed configural measurement invariance with RMSEA and CFI, indicating a very good fit (see Table 3.5). Therefore, a meaningful discussion of constructive patriotism and nationalism across countries was possible (Davidov et al. 2014).

Given the reassuring values from the configural invariance test, we moved on to test for metric invariance. This test also requires equal factor loadings (Rock, Werts, and Flaugher 1978). As the sample size affects the chi-square difference test, we instead compared the two models with the criteria that Chen (2007) proposed. The Δ CFI did not exceed .01, and the Δ RMSEA was below .015. Therefore, metric invariance could be established, and exploring the structural relationships with other constructs across countries was possible (Steenkamp and Baumgartner 1998).

Table 3.5. MGCFA: Fit Measures of the Measurement Invariance Tests

<i>Model</i>	χ^2	<i>df</i>	$\Delta RMSEA$	<i>RMSEA</i>	ΔCFI	<i>CFI</i>
1. Configural	112.869	20		.061 [.051; .073]		.982
2. Full metric	144.424	32	-.008	.053 [.045; .062]	-.003	.979
3. Scalar invariance	1582.351	42	+.116	.169 [.162; .176]	-.269	.710
3a. Partial scalar: [V25]	830.976	40	+.074	.127 [.119; .134]	-.128	.851
3b. Partial scalar: [V28]	742.845	40	+.067	.120 [.112; .127]	-.112	.867
3c. Partial scalar: [V34]	1265.001	40	+.105	.158 [.150; .165]	-.210	.769

We continued to test for scalar invariance, which is a precondition for a cross-national comparison of the constructs' mean values. The test additionally asks for equal intercepts (Meredith 1993). As ΔCFI and $\Delta RMSEA$ clearly exceeded their critical values ($\Delta RMSEA$: .116, ΔCFI : .269), full scalar invariance could not be established. We still tested for partial scalar invariance (Steenkamp and Baumgartner 1998), which requires at least two items with equal intercepts. As nationalism is measured by only two indicators, a partial scalar invariance test is only viable for constructive patriotism (measured by three indicators). We applied the same research strategy as Davidov (2009) and estimated three additional models, in which we separately freed the intercepts of one of the indicators of constructive patriotism for all countries. Although the GOF indices improved, especially when freeing the intercept for the "social security" item, they did not improve enough to establish partial scalar invariance. Hence, a cross-national comparison of the means of the latent constructs is impossible, but exploring the structural relationships with other constructs across these five countries is possible. We turn now to the OP results for the three items that measure constructive patriotism: "democracy," "social security," and "fair and equal." We will limit our description to the most relevant substantive findings and we will focus on the results that help explain why the (partial) scalar invariance failed in MGCFA.

3.5 RESULTS: ONLINE PROBING

3.5.1 Category-selection Probe for “Democracy”

Although cross-national studies about democracy abound, several studies indicate that respondents conceptualize democracy in different ways (Baviskar and Malone 2004; Canache et al. 2001). However, Behr and Braun (2015) note that the very fact that respondents in all countries have a similar multiplicity of democratic concepts makes this item cross-nationally comparable. For this study, we used an adapted version of Behr and Braun’s coding schema. For us, respondents’ acceptance of the democratic system is an important precondition for the cross-national comparability of the “democracy” item. Low pride values should reflect that the system does not live up to the respondents’ expectations; it should not reflect the respondents’ preferences for a more authoritarian form of government.

The coding schema for the “democracy” item. The coding schema for the category-selection probe captures positive and negative evaluations of different aspects of democracy. Some respondents thought about the *output of democratic authorities*, which might have entailed an evaluation of the living conditions in a country or specific policy domains (e.g., positive: tolerant society, nuclear phase-out in Germany; negative: growing insecurity, tax increases). Other respondents evaluated whether the *governance, politicians or other authorities*, were working according to democratic ideals and rules (e.g., positive: low level of lobbying and corruption; negative: high level of lobbying and corruption, politicians’ poor character). The respondents also perceived the *political system, the institutions or the constitutional arrangements* as (un)democratic (e.g., democratic: free elections, working rule of law; undemocratic: malfunctioning of checks and balances, no freedom of speech). The respondents who considered other political systems superior to the democratic system would be coded here. A few respondents disapproved of the *lack of citizens’ support in upholding democracy* (e.g., low voter turnout) or evaluated the democratic situation in a more general

sense. They mentioned that a *working democracy existed*, that the democracy *needed some improvement* or that *no democracy existed*. Some respondents also *compared their country with other countries*. Other respondents were proud of their country *independent of its democratic situation*, and some identified *problems with the question*—two potentially problematic codes for cross-national comparability. All remaining answers were coded as *rest*.

Probe results for the “democracy” item. Regarding the *output of authorities*, most respondents in all countries focused on negative aspects, mostly complaining about the growing inequality in their society or about recent government policies (see Table 3.6). Mexican participants differ from the other respondents because they refer to the problematic security situation in their country (e.g., “because of the existing criminality”), an issue that respondents from the other countries did not mention. This difference is reflected in CFA results of the Mexican single-country analysis, which suggested an error correlation between the “democracy” and “social security benefits” items (see 3.5.2 for further details).

Table 3.6. Proportion of Codes for the “Democracy” Item (Substantive Responses)

Code	Germany (N = 94) (%)	GB (N = 100) (%)	U.S. (N = 96) (%)	Mexico (N = 117) (%)	Spain (N = 110) (%)
1P: Output authorities: Positive evaluation	1	1	4	0	0
1N: Output authorities: Negative evaluation	12	11	7	12	12
2P: Governance: Positive evaluation	1	1	1	1	0
2N: Governance: Negative evaluation	18	17	16	34	45
3P: Political system, institutions, and constitutional arrangements perceived as democratic	14	13	13	3	8
3N: Political system, institutions, and constitutional arrangements perceived as undemocratic	15	18	12	28	28
4N: Lack of citizens’ support in upholding democracy	3	7	2	1	3
5: A working democracy exists	15	16	7	0	1
6: Democracy can be improved	4	10	8	5	7
7: No democracy exists	9	4	4	20	11
8: Comparison with other countries	11	15	15	2	5
9: Pride judgment independent of democratic situation	0	3	6	0	0
10: Problems with the question	2	5	1	1	1
11: Rest	14	11	13	8	6

Although respondents from all countries mostly evaluated the *governance* aspect of democracy negatively (Germans: lobbying; other countries: corrupt politicians), a higher proportion of Spaniards and Mexicans mentioned negative aspects compared with those in other countries. This response pattern reflects the diverging realities in the different countries (e.g., the higher level of corruption in Mexico; Freedom House 2015), but it does not necessarily threaten cross-national comparability.

The cross-national differences also hold for the evaluation of the general set-up of the *political system, institutions, and constitutional arrangements*. Unlike the previous dimensions, the respondents associated positive aspects, such as freedom of speech (all countries) and free elections (all but Mexico), with this democratic aspect. Meanwhile, several respondents perceived the general set-up of the democratic system as undemocratic (e.g., Germany: lack of direct democracy; Great Britain: first-past-the-post voting; U.S.: discontent with Congress). Most Spanish and Mexican respondents mentioned negative aspects in this context. Spanish respondents disapproved of the insufficient separation of powers, the lack of direct participation and the two-party system in their country. Mexican respondents complained more severely about the general set-up of the democratic system, highlighting the failure of the judiciary and the lack of freedom of speech and free elections. This evaluation reflects the current state of Mexico's democracy, which was downgraded from "free" to "partly free" in Freedom House's 2014 annual report (Freedom House 2015). Interestingly, none of the respondents rejected the idea of democracy altogether and indicated that they would prefer a more authoritarian system instead. This finding is reassuring because a condition for this item's cross-national comparability is that low pride values do not simultaneously reflect discontent with the democratic system and an endorsement of more authoritarian alternatives.

Two other codes had the potential to threaten the cross-national comparability of the “democracy” item: *Pride judgment independent of the democratic situation* and *problems with the question*. Fortunately, only a few U.S. respondents mentioned the former, and German and British respondents rarely mentioned problems with the question. Given the low percentages in both categories, they do not pose a threat to comparability.

Overall, the probing results reveal many different perspectives that respondents adopt when they answer this item. Since respondents in all countries think about various aspects of democracy, this item is still cross-nationally comparable. The absence of respondents who prefer authoritarian rule rather than a democratic system and the low proportion of respondents with pride judgments independent of the democratic situation and problems with the question are reassuring.

3.5.2 Specific Probe for “Social Security”

The coding schema for the “social security” item. We developed a second coding schema for the answers to the specific probe for the “social security” item. Many respondents mentioned *welfare benefits* that provide a minimal level of well-being and social support to all citizens (e.g., the U.S.: food stamps; Great Britain: housing benefits). The item also triggered references to *unemployment benefits* (e.g., jobseeker allowance, specific training courses), *health-related benefits* (e.g., health insurance, disability benefits) or *retirement benefits*. *Family benefits* (e.g., Germany: maternity/paternity leave; Great Britain: child tax credits) and *support for immigrants or refugees* were also mentioned. Unexpectedly, Mexican respondents referred to the *security* situation in their country (e.g., too much violence and crime). Several respondents referred to *all benefits* or wrote *ambiguous answers* that mentioned agencies that deliver more than one benefit (e.g., Mexico: IMSS) without explaining further. All remaining answers were coded as *rest*.

Probe results for the “social security” item. Overall, German respondents and British respondents mentioned a wider variety of social security benefits than U.S., Mexican, and Spanish respondents (see Table 3.7). For example, German and British respondents more frequently wrote *unemployment benefits*, *family benefits* and *support for immigrants and refugees* than participants from other countries. Most U.S. respondents (68 percent) referred to *retirement benefits*, and 72 percent of Spanish respondents mentioned *health-related benefits*. Given the different types of welfare states in the five countries, it might not be surprising that respondents associated different benefits with their social security system. Two critical points remain that give indications of other influences that might affect the response behavior of Americans, Spaniards, and Mexicans.

Table 3.7. Proportions of Codes for the “Social Security” Item (Substantive Responses)

Code	Germany (N = 75) (%)	GB (N = 64) (%)	U.S. (N = 71) (%)	Mexico (N = 96) (%)	Spain (N = 85) (%)
Welfare benefits	28	19	7	4	0
Unemployment benefits	51	27	3	4	5
Health-related benefits	42	38	21	39	72
Retirement benefits	25	11	68	4	11
Family benefits	23	9	0	1	4
Support for immigrants and refugees	7	11	0	0	2
Security	0	0	0	39	0
All benefits	3	14	7	8	9
Rest	19	12	3	6	5
Ambiguous answers	1	11	7	21	19

First, the range of perceived benefits varies across countries because of the translation of the term “social security benefits.” For example, the U.S. system offers more than retirement benefits, including Medicaid as a health benefit and food stamps and public housing as welfare benefits. The closed item (which was also used in the ISSP study) asked, “How proud are you of America with regard to its *social security system*?” In the U.S., the agency responsible for retirement benefits is called the *Social Security Administration*, and *social security* taxes contribute to the retirement system. In many probe responses that were coded as retirement benefits, U.S. respondents made no distinction between general social

security benefits and retirement benefits. The following probe answers from two U.S. respondents exemplify the lack of distinction between the two:

Social security administration. Money American's are supposed to receive when they retire. (American, "somewhat proud")

Social security for seniors. (American, "somewhat proud")

The Spanish system also provides several social security benefits, but many respondents were only thinking about health care. For some respondents, "seguridad social," the Spanish translation of "social security benefits," is seemingly the equivalent of the health care system in Spain. In the following probe answer, the respondent used "seguridad social" interchangeably with the health care system:

The basic benefits. There are not enough doctors ... in the hospitals. A lot of specialists and drugs are not covered by the social security system [la seguridad social]. The general practitioners are in most cases just graduates who do not have a clue. (Spaniard, "not proud at all")

These translations pose a problem for comparability because in the German version the term "social security benefits" was translated as "sozialstaatliche Leistungen," which can refer to any kind of social security benefits, such as unemployment benefits, health insurance, and family benefits. Therefore, the German respondents were answering a question that had a larger lexical scope than that of the U.S. and Spanish respondents.

The varying range of perceived social security benefits does have an impact on this item's cross-national comparability. As the MGCFA results have shown, a cross-national comparison of the latent means of nationalism and constructive patriotism is impossible. To achieve partial scalar invariance, we freed the intercepts of each item that measured constructive patriotism. Although we could not achieve partial scalar invariance, freeing the intercepts for the "social security" item would have yielded the greatest model improvement, as the ΔCFI values were smaller than those in the solution when the intercept was freed for the "fair and equal" and "democracy" items ("social security": ΔCFI : -.112; "democracy":

-.128; “fair and equal”: -.210); therefore, the “social security” item is potentially the most problematic item in a cross-national comparison. The varying lexical scope might partially explain this outcome.

The second issue regarding the “social security” item is the high proportion of Mexican respondents (39 percent) who mentioned the general security situation in Mexico instead of its social security system. Mexican probe responses show that the respondents were thinking about violence, crime, or robberies. The second respondent even noted that he was uncertain about the intended meaning of the question:

Is there such a thing? Crime, drug trafficking, attacks, robberies, police brutality, etc. Social security? That’s a myth. (Mexican, “not proud at all”)

Which social security? If we speak about crimes, it gets worse each day, and if we speak about health insurance, the service is bad, inefficient; they make fun of us. (Mexican, “not proud at all”)

Given the problematic security situation in Mexico, these respondents might be more inclined than respondents from other countries to perceive the increased violence and criminality of their surroundings. For example, between 2006 and 2011, Mexico registered 47,515 crime-related deaths (Gonzalez 2012). A substantive part of the Mexican respondents misunderstood the intended meaning of the question, which again reduces this item’s cross-national comparability. Interestingly, the nonresponse rate of the closed item was particularly low for Mexican respondents. These responses are a classic example of “silent misinterpretation,” where respondents are unaware that they have misunderstood the item’s intended meaning (DeMaio and Rothgeb 1996).

Silent misinterpretation can also explain one of the MGCFA findings. In the Mexican single-country analysis, Jrule suggested allowing for an error correlation between the “democracy” and “social security” items. As we have seen, probe responses for the “democracy” item revealed that some Mexicans worry about the security situation in their country. On the “social security” item, some Mexican respondents also expressed their

concern about the security situation in Mexico. In addition to the constructive patriotism factor, the concern about the general security situation is a factor that influences the variance between both items, which might explain the correlation between the error terms for the “democracy” and “social security” items.

The probe results for the “social security” item elucidate several problematic issues for cross-national comparability. The varying lexical scope of the term “social security system” and its silent misinterpretation by several Mexican respondents can potentially explain why the scalar measurement invariance tests failed.

3.5.3 *Specific Probe for “Fair and Equal”*

Finally, after the “fair and equal” item, we asked a specific probe about the social group that the respondents had in mind. Although previous articles have considered the vague formulation of the term “social groups in society” problematic (Fleiß et al. 2009; Latcheva 2011; Meitinger and Behr, forthcoming), the “fair and equal” item was the least problematic indicator for the measurement invariance tests. Freeing the item’s intercept would have yielded the smallest model improvement when testing for partial scalar invariance (“social security”: $\Delta CFI: -.112$; “democracy”: $-.128$; “fair and equal”: $-.210$). However, for the U.S., Jrule suggested an additional cross loading on the nationalism factor in the single-country analysis. We accept that respondents adopt different perspectives when they answer this item, since the social realities vary in the five countries. Most importantly, the answer should reflect the respondent’s stance on constructive patriotism, particularly the value of equality.

The coding schema for the “fair and equal” item. The coding schema distinguishes between *foreigners* in an abstract sense (e.g., “immigrants”), *specific nationalities* (e.g., “Indian”), and *specific races or ethnicities* (e.g., “Native Americans,” “Hispanics”). The respondents also mentioned the *vertical division* in society (e.g., “the rich and the poor”),

religion (e.g., “Muslims”), *gender*, and *sexuality* (e.g., “homosexuals”). Some participants thought about *ill people* and *older citizens*. All these groups can be considered potential targets of mistreatment. However, the respondents also switched perspectives and referred to *groups that are perceived as having a detrimental impact on society*, such as politicians, bankers, and judges. Furthermore, some respondents associated this term with the *majority group*. This code is a potential indicator that the item does not serve as a good indicator for constructive patriotism. Instead, it might reflect nationalistic attitudes when respondents show in-group favoritism (e.g., when they complain about an overabundance of help for foreigners and fear that the majority group’s benefits are being reduced). These statements often used derogatory language when referring to foreigners. If a high proportion of U.S. respondents mentioned this code, the CFA finding suggesting an additional cross loading of this item on the nationalism factor would be confirmed for the U.S. Additionally, several respondents referred to either *all groups* or *no specific group*. All remaining answers were coded as *rest*.

Probe results for the “fair and equal” item. In all countries, the respondents think about a wide variety of different groups (see Table 3.8). However, the countries differ in terms of the groups that come to mind most. For example, German respondents more often mentioned *foreigners* in general, whereas respondents in the U.S. thought more frequently about different *races or ethnicities*. Mexican and Spanish respondents were particularly concerned about the vertical division in their countries (e.g., the poor versus the upper class).

Some respondents switched their perspectives when they answered the question about those *responsible for unfair and unequal treatment* (e.g., politicians, bankers or judges), particularly in Spain (35 percent). However, this change in perspective does not constitute a threat to cross-national comparability when the item is used as an indicator of constructive patriotism. The respondents who mentioned this category were still concerned about democratic values.

Table 3.8. Proportions of Codes for the “Fair and Equal” Item (Substantive Responses)

Code	Germany (N = 89) (%)	GB (N = 94) (%)	U.S. (N = 87) (%)	Mexico (N = 84) (%)	Spain (N = 99) (%)
Foreigners	37	18	2	0	29
Specific nationalities	8	12	11	0	12
Race and ethnicity	6	34	56	37	11
Vertical division: Top – bottom	56	26	14	71	60
Religion	9	27	14	1	6
Gender	11	13	9	13	25
Sexual orientation	22	20	18	6	11
Ill people	17	26	3	5	3
Older citizens	20	9	0	8	3
Groups exerting a detrimental impact on society	12	6	6	18	35
Majority	3	6	1	1	12
All groups	6	9	11	10	4
No specific group	1	3	13	0	1
Rest	27	19	13	29	21

The respondents who referred to the *majority group* in their country are potentially more problematic for cross-national comparability. This code could reflect nationalistic values rather than patriotic values. Fortunately, the proportion of respondents who mention this category is too low for concern in Germany, Great Britain, and Mexico. Interestingly, only one U.S. respondent referred to this category, which contradicts the CFA results for the U.S. single-country analysis. Jrule indicated a misspecification for the U.S. in this item and recommended an additional loading on the nationalism factor. The OP results did not support the initial speculation that Americans felt superior to other countries because of their country’s focus on egalitarian issues. In addition, we did not find any indications of American respondents’ nationalistic attitudes at the category-selection probe that we also used for this item⁷. The OP results confirm our decision to refrain from a respecification of the U.S. model. The decision to modify a model should never be driven by pure reliance on MIs and EPCs; it should instead be guided by substantive theory (Brown 2006). OP can also help discern between cases in which a model specification would have been appropriate (Mexico: error correlation) and those in which it would have been inappropriate (as in the U.S.). In contrast to the other nationalities, Spaniards more often referred to the majority (12 percent). However,

⁷ The detailed results and the full coding schema are available from the author upon request.

seven of the eleven respondents in this category mentioned the majority group when complaining about those responsible (e.g., banker and politicians) for the poor economic situation in their country: “About politicians and bankers, it does not seem that they care about the judiciary—about the majority of society, who has to deal with eviction notices” (Spaniard, “not very proud”).

Despite multiple associations and perspectives, the “fair and equal” item still seems cross-nationally comparable. The differences in the mentioned social groups reflect the complex social reality in the five countries. The OP results provide reassurance that the item serves as a good indicator of egalitarian attitudes, which was also reflected in the tests for partial scalar variance. However, the OP results did not support the suggested model modification in the CFA single-country analysis for the U.S.

3.6 DISCUSSION

This study sought to compare the MGCFA and OP methods. Using the example of constructive patriotism, we wanted to test whether these methods would arrive at similar conclusions. The MGCFA was able to confirm metric invariance, but (partial) scalar invariance tests failed. Exploring structural relationships is possible, but a cross-national comparison of the means of the latent constructs is impossible. In addition, GOF indices and the MIs and EPCs in Jrule suggested two model modifications in the single-country analysis. For the U.S., Jrule recommended letting the “fair and equal” item also load onto the nationalism factor. For Mexico, Jrule advised allowing for an error correlation between the “democracy” and “social security” items.

Indeed, the OP results did partly clarify the MGCFA results. In particular, the probe for the “social security” item uncovered several problematic issues that were mirrored in MGCFA findings. First, OP explained that the suggested error correlation between the “democracy” and “social security” items in Mexico was driven by a silent misinterpretation of the term

“social security system.” Many Mexican respondents understood “security” instead of “social security.” The Mexican security situation was also an issue for the “democracy” item. Therefore, concerns about general security affected part of the correlation between the two items in Mexico. Second, MGCFA indicated that freeing the intercepts of the “social security” item would yield the largest model improvements. OP revealed that the range of perceived benefits varied across countries. The different translations triggered either references to specific benefits, such as “retirement” in the U.S. and “health” in Spain, or to a wide variety of possible benefits (Germany). These country-specific components explain why the test for scalar invariance failed and why means should not be compared across countries.

By contrast, OP could not confirm the initial speculation on the “fair and equal” item in the U.S. Jrule suggested letting this item load onto the nationalism factor. Neither the specific probe nor the category-selection probe was able to validate the initial hypothesis that American nationalists felt superior to other countries because of their country’s focus on egalitarian issues. OP is a handy tool for evaluating the appropriateness of such ad hoc hypotheses and a guide for model modifications.

The previous results show that much can be learned through a combined approach of MGCFA and OP. Depending on the research situation, two strategies for combining these methods are proposed. In an *exploratory* research situation, developing new items is necessary. OP could be implemented in the pretest stage to guide the cross-national item development. As OP is limited to a small number of countries, the developed items could be tested in a second step with a larger number of countries using MGCFA. By contrast, in a research situation in which more *established* survey items are used with a large number of countries, first conducting the different measurement invariance tests with MGCFA might be advisable. If some of the tests detect noninvariance, MIs and EPCs could help find the items and countries that should be used in an OP study. The probes can elucidate the reasons for the lacking cross-national comparability of these items or countries.

References

- Adorno, Theodor W., Else Frenkel-Brunswik, Daniel J. Levinson, and R. Nevitt Sanford. 1950. *The Authoritarian Personality*. Oxford: Harpers.
- Baviskar, Siddhartha and Mary Fran T. Malone. 2004. "What Democracy Means to Citizens—and Why It Matters." *Revista Europea de Estudios Latinoamericanos y del Caribe/European Review of Latin American and Caribbean Studies* 76:3–23.
- Beatty, Paul C. and Gordon B. Willis. 2007. "Research Synthesis: The Practice of Cognitive Interviewing." *Public Opinion Quarterly* 71(2):287–311.
- Behr, Dorothee and Michael Braun. 2015. "Satisfaction with the Way Democracy Works: How Respondents across Countries Understand the Question." Pp. 121–138 in *Hopes and Anxieties. Six Waves of the European Social Survey*, edited by P. B. Sztabinski, H. Domanski, and F. Sztabinski. Frankfurt am Main: Lang.
- Bentler, Peter M. 1990. "Comparative Fit Indexes in Structural Models." *Psychological Bulletin* 107(2):238–46.
- Blank, Thomas, Peter Schmidt, and Bettina Westle. 2001. "Patriotism—A Contradiction, a Possibility or an Empirical Reality." in *European Consortium for Political Research Joint Sessions of Workshops*, Grenoble, France.
- Blank, Thomas and Peter Schmidt. 2003. "National Identity in a United Germany: Nationalism or Patriotism? An Empirical Test with Representative Data." *Political Psychology* 24:289–311.
- Braun, Michael, Dorothee Behr, Lars Kaczmarek, and Wolfgang Bandilla. 2014. "Evaluating Cross-national Item Equivalence with Probing Questions in Web Surveys." Pp. 184–200 in *Improving Survey Methods: Lessons from Recent Research*, edited by U. Engel, B. Jann, P. Lynn, A. Scherpenzeel, and P. Sturgis. New York: Routledge.

- Braun, Michael and Timothy Johnson. 2010. "An Illustrative Review of Techniques for Detecting Inequivalences." Pp. 373–93 in *Survey Methods in Multinational, Multiregional, and Multicultural Contexts*, edited by J. A. Harkness, M. Braun, B. Edwards, T. P. Johnson, L. Lyberg, P. Ph. Mohler, B. Pennell, and T. W. Smith. Hoboken, NJ: Wiley-Blackwell.
- Brown, Timothy A. 2014. *Confirmatory Factor Analysis for Applied Research*. New York: Guilford Publications.
- Browne, Michael W. and Robert Cudeck. 1992. "Alternative Ways of Assessing Model Fit." *Sociological Methods & Research* 21(2):230–58.
- Byrne, Barbara M., Richard J. Shavelson, and Bengt Muthén. 1989. "Testing for the Equivalence of Factor Covariance and Mean Structures: The Issue of Partial Measurement Invariance." *Psychological Bulletin* 105(3):456–66.
- Byrne, Barbara M. and Fons J.R. van De Vijver. 2010. "Testing for Measurement and Structural Equivalence in Large-scale Cross-cultural Studies: Addressing the Issue of Nonequivalence." *International Journal of Testing* 10(2):107–32.
- Canache, Damary, Jeffery J. Mondak, and Mitchell A. Seligson. 2001. "Meaning and Measurement in Cross-national Research on Satisfaction with Democracy." *Public Opinion Quarterly* 65(4):506–28.
- Chen, Fang Fang. 2007. "Sensitivity of Goodness of Fit Indexes to Lack of Measurement Invariance." *Structural Equation Modeling* 14(3):464–504.
- Cheung, Gordon W. and Roger B. Rensvold. 2002. "Evaluating Goodness-of-fit Indexes for Testing Measurement Invariance." *Structural Equation Modeling* 9(2):233–55.
- Cieciuch, Jan and Eldad Davidov. (forthcoming). "Establishing Measurement Invariance across Online and Offline Samples. A Tutorial with the Software Packages Amos and Mplus." *Studia Psychologica*.

- Davidov, Eldad. 2009. "Measurement Equivalence of Nationalism and Constructive Patriotism in the ISSP: 34 Countries in a Comparative Perspective." *Political Analysis* 17(1):64–82.
- Davidov, Eldad. 2011. "Nationalism and Constructive Patriotism: A Longitudinal Test of Comparability in 22 Countries with the ISSP." *International Journal of Public Opinion Research* 23(1):88–103.
- Davidov, Eldad, Georg Datler, Peter Schmidt, and Shalom H. Schwartz. 2011. "Testing the Invariance of Values in the Benelux Countries with the European Social Survey: Accounting for Ordinality." Pp. 149–68 in *Cross-cultural Analysis: Methods and Applications*, edited by E. Davidov, P. Schmidt, and J. Billiet. New York: Routledge.
- Davidov, Eldad, Hermann Dülmer, Elmar Schlüter, Peter Schmidt, and Bart Meuleman. 2012. "Using a Multilevel Structural Equation Modeling Approach to Explain Cross-cultural Measurement Noninvariance." *Journal of Cross-cultural Psychology* 43(4):558–75.
- Davidov, Eldad, Bart Meuleman, Jan Cieciuch, Peter Schmidt, and Jaak Billiet. 2014. "Measurement Equivalence in Cross-national Research." *Annual Review of Sociology* 40:55–75.
- De Beuckelaer, Alain and Gilbert Swinnen. 2011. "Biased Latent Variable Mean Comparisons Due to Measurement Noninvariance: A Simulation Study." Pp. 117–47 in *Cross-cultural Analysis: Methods and Applications*, edited by E. Davidov, P. Schmidt, and J. Billiet. New York: Routledge.
- DeMaio, Theresa J. and Jennifer M. Rothgeb. 1996. "Cognitive Interviewing Techniques: In the Lab and in the Field." Pp. 177–95 in *Answering Questions: Methodology for Determining Cognitive and Communicative Processes in Survey Research*, edited by N. Schwarz and S. Sudman. San Francisco: Jossey-Bass.

- Fleiß, Jürgen, Franz Höllinger, and Helmut Kuzmics. 2009. "Nationalstolz zwischen Patriotismus und Nationalismus?" *Berliner Journal für Soziologie* 19(3):409–34.
- Flora, David B. and Patrick J. Curran. 2004. "An Empirical Evaluation of Alternative Methods of Estimation for Confirmatory Factor Analysis with Ordinal Data." *Psychological Methods* 9(4):466–91.
- Freedom House. 2015. *Freedom in the World 2014: The Annual Survey of Political Rights and Civil Liberties*. New York: Rowman & Littlefield.
- Gray, Michelle and Margaret Blake. 2015. "Cross-national, Cross-cultural and Multilingual Cognitive Interviewing." Pp. 220–42 in *Cognitive Interviewing Practice*, edited by Debbie Collins. London: Sage Publication.
- Gonzalez, Francisco. 2012. *Freedom House – Countries at the Crossroads 2012: Mexico*. Washington D.C.: Freedom House. Retrieved May 20, 2015 (https://freedomhouse.org/report/countries-crossroads/2012/mexico#.VWhkQmM0_jA).
- Horn, John L. and Jack J. McArdle. 1992. "A Practical and Theoretical Guide to Measurement Invariance in Aging Research." *Experimental Aging Research* 18(3):117–44.
- Hu, Li-tze and Peter M. Bentler. 1999. "Cutoff Criteria for Fit Indexes in Covariance Structure Analysis: Conventional Criteria versus New Alternatives." *Structural Equation Modeling: A Multidisciplinary Journal* 6(1):1–55.
- Hui, C. Harry and Harry C. Triandis. 1985. "Measurement in Cross-cultural Psychology: A Review and Comparison of Strategies." *Journal of Cross-cultural Psychology* 16(2):131–52.
- ISSP Research Group. 2015. *International Social Survey Programme: National Identity III - ISSP 2013. ZA5950 Data file Version 1.0.0*. Cologne: GESIS Data Archive. doi: 10.4232/1.12195.

- Jöreskog, Karl G. 1971. "Simultaneous Factor Analysis in Several Populations." *Psychometrika* 36(4):409–26.
- Kosterman, Rick and Seymour Feshbach. 1989. "Toward a Measure of Patriotic and Nationalistic Attitudes." *Political Psychology* 10(2):257–74.
- Latcheva, Rossalina. 2011. "Cognitive Interviewing and Factor-analytic Techniques: A Mixed Method Approach to Validity of Survey Items Measuring National Identity." *Quality & Quantity* 45(6):1175–99.
- Meitinger, Katharina and Dorothée Behr. (forthcoming). "Comparing Cognitive Interviewing and Online Probing: Do They Find Similar Results?" *Field Methods*.
- Meredith, William. 1993. "Measurement Invariance, Factor Analysis and Factorial Invariance." *Psychometrika* 58(4):525–43.
- Meuleman, Bart. 2012. "When Are Item Intercept Differences Substantively Relevant in Measurement Invariance Testing?" Pp. 97–104 in *Methods, Theories, and Empirical Applications in the Social Sciences*, edited by S. Salzborn, E. Davidov, and J. Reinecke. Wiesbaden: VS Verlag für Sozialwissenschaften.
- Meuleman, Bart and Jaak Billiet. 2009. "A Monte Carlo Sample Size Study: How Many Countries are Needed for Accurate Multilevel SEM?" *Survey Research Methods* 3(1):45–58.
- Miller, Kirsten, Daniel Mont, Aaron Maitland, Barbara Altman, and Jennifer Madans. 2011. "Results of a Cross-national Structured Cognitive Interviewing Protocol to Test Measures of Disability." *Quality & Quantity* 45(4):801–15.
- Oberski, Daniel. 2014. *Jrule for Mplus: A Program for Post-hoc Power Evaluation of Structural Equation Models Estimated by Mplus*. doi: 10.5281/zenodo.10657.

- Prüfer, Peter and Margrit Rexroth. 2005. "Kognitive Interviews." *ZUMA How-to-Reihe 15*.
Retrieved October 20, 2015
(http://www.gesis.org/fileadmin/upload/forschung/publikationen/gesis_reihen/howto/How_to15PP_MR.pdf).
- Rock, Donald A., Charles E. Werts, and Ronald L. Flaughner. 1978. "The Use of Analysis of Covariance Structures for Comparing the Psychometric Properties of Multiple Variables across Populations." *Multivariate Behavioral Research* 13(4):403–18.
- Saris, Willem E., Albert Satorra, and William M. van der Veld. 2009. "Testing Structural Equation Models or Detection of Misspecifications?" *Structural Equation Modeling* 16(4):561–82.
- Schatz, Robert T., Ervin Staub, and Howard Lavine. 1999. "On the Varieties of National Attachment: Blind versus Constructive Patriotism." *Political Psychology* 20(1):151–74.
- Steenkamp, Jan-Benedict E. and Hans Baumgartner. 1998. "Assessing Measurement Invariance in Cross-national Consumer Research." *Journal of Consumer Research* 25(1):78–107.
- Van de Vijver, Fons J. 2011. "Capturing Bias in Structural Equation Modeling." Pp. 3–34 in *Cross-cultural Analysis. Methods, Theories, and Empirical Applications in the Social Sciences*, edited by S. Salzborn, E. Davidov, and J. Reinecke. Wiesbaden: VS Verlag für Sozialwissenschaften
- Vandenberg, Robert J. and Charles E. Lance. 2000. "A Review and Synthesis of the Measurement Invariance Literature: Suggestions, Practices, and Recommendations for Organizational Research." *Organizational Research Methods* 3(1):4–70.
- Willis, Gordon B. 2005. *Cognitive Interviewing: A Tool for Improving Questionnaire Design*. Thousand Oaks, CA: Sage.

Appendix

RESULTS WITH THE WLSMV APPROACH: ACCOUNTING FOR ORDINALITY

Table S3.1. Single-country Analyses with WLSMV Estimator: RMSEA, CFI, and Correlations between Nationalism and Constructive Patriotism and Error Correlations (Standard Errors in Parentheses)

<i>Country</i>	<i>RMSEA</i>	<i>CFI</i>	<i>Correlation</i>
1. Germany	.075 [.055;.096]	.988	N↔CP: .38 (.03)
2. Great Britain	.000 [.000;.046]	1.000	N↔CP: .39 (.04)
3. Mexico	.091 [.066;.118]	.989	N↔CP: .57 (.03)
4. Spain	.067 [.044;.093]	.991	N↔CP: .67 (.03)
5. U.S.	.087 [.065;.112]	.961	N↔CP: .50 (.04)

Table S3.2. Single-country Analysis with WLSMV Estimator: Unstandardized and Standardized Factor Loadings and Standard Errors

<i>Country</i>	<i>N</i> → <i>V19</i>	<i>N</i> → <i>V20</i>	<i>CP</i> → <i>V25</i>	<i>CP</i> → <i>V28</i>	<i>CP</i> → <i>V34</i>
(a) Factor loadings on nationalism and constructive patriotism (unstandardized) (standard error in parentheses)					
1. Germany	.75 (.07)	1	1	.94 (.05)	.93 (.05)
2. Great Britain	.68 (.09)	1	1	.89 (.06)	.99 (.08)
3. Mexico	.88 (.06)	1	1	1.00 (.04)	.89 (.03)
4. Spain	.76 (.04)	1	1	.72 (.05)	.87 (.05)
5. U.S.	.90 (.11)	1	1	.92 (.08)	1.17 (.11)
(b) Factor loadings on nationalism and constructive patriotism (standardized) (standard error in parentheses)					
1. Germany	.70 (.03)	.94 (.04)	.73 (.02)	.69 (.03)	.68 (.02)
2. Great Britain	.63 (.05)	.93 (.06)	.70 (.04)	.62 (.03)	.69 (.03)
3. Mexico	.73 (.03)	.83 (.03)	.81 (.02)	.81 (.02)	.72 (.02)
4. Spain	.69 (.02)	.90 (.03)	.76 (.03)	.55 (.03)	.66 (.03)
5. U.S.	.65 (.04)	.72 (.05)	.57 (.04)	.53 (.03)	.67 (.04)

Note: N = nationalism factor; CP = constructive patriotism factor.

Table S3.3. MGCFA with WLSMV Estimator: Fit Measures of the Measurement Invariance Test

<i>Model</i>	<i>df</i>	<i>RMSEA</i>	<i>CFI</i>
1. Configural	20	.074 [.063;.085]	.988
2. Scalar invariance	72	.165 [.160;.171]	.787

Note: It is necessary to constrain factor loading, intercepts, and thresholds when testing ordinal data for measurement invariance. Therefore, we refrain from reporting the results for the metric invariance test (see also Davidov et al. 2011).

4. What Does the General National Pride Item Measure?: Insights from Online Probing⁸

Abstract

The general national pride question is a popular item in cross-national studies of national identity. Although it is a single-item indicator, researchers use it as a proxy for complex concepts, such as patriotism and nationalism. This article assesses the suitability of the general national pride item for cross-national studies. By means of an exploratory factor analysis and online probing results from a web survey conducted in five countries, we reveal that the general national pride item is a problematic indicator for cross-national studies of national identity and its elements. In addition to the cross-national variability of respondents' associations, the probe uncovered several problematic issues that had a distorting impact on the answer selection. These results caution against the use of single-item indicators in a cross-national context but strongly call for a measurement with multiple indicators that enable an assessment of cross-national comparability.

⁸ Preliminary results were presented at the ESRA 2015: 6th Conference of the European Survey Research Association, July 13-17, 2015, Reykjavik, Iceland.

4.1 INTRODUCTION

The study of national identity is an important research field that has gained popularity in recent years. For example, the current bibliography of the International Social Survey Program (ISSP) already lists 512 entries for the Module on National Identity (Smith and Schapiro 2015) and the arrival of the 2013 Module on National Identity (ISSP Research Group 2015) will clearly increase this number. Apart from the ISSP, several other large scale cross-national studies include questions about national identity (e.g., the World Value Survey [WVS]), but the ISSP is unique in the amount of questions it asks about this topic.

Despite the growing significance of this research area, criticism regarding the studies of national identity persists. Malešević (2011) even classifies national identity as a “conceptual monstrosity”: “The concept of ‘national identity’ is a sweeping conceptual chimera that is often simply used as a description of assumed social reality or invoked as a shortcut explanation for particular forms of collective behaviour” (p. 281).

In a similar vein, several researchers have complained about the lack of consensus on theoretical definitions regarding *national identity* (Davidov 2009; Fleiß, Höllinger, and Kuzmics 2009; Latcheva 2011). Indeed, national identity can mean various things to different researchers. Some perceive it as a single dimension (e.g., Elkins and Sides 2006), but most researchers dichotomize national identity. On the one hand, they differentiate between a civic and an ethnic form of national identity (e.g., Reeskens and Wright 2013; Smith 1991). This approach often studies which criteria are used to distinguish between who is perceived as belonging to a nation and who does not. On the other hand, some researchers make a distinction between patriotism and nationalism (Adorno et al. 1950; Blank and Schmidt 2003; Kosterman and Feshbach 1989; Schatz, Staub, and Lavine 1999), two expressions of national affection (Latcheva 2011). Even more problematic, the field also has struggled to find a consensus on the definitions of the different elements of national identity, such as *patriotism*:

Research on patriotism has been marred by a confusing array of terms, definitions, and expected consequences in which patriotism is variously defined as a sense of national loyalty, a love of national symbols, specific beliefs about a country's superiority, and as a crucial ingredient in the development of civic ties to a mature nation. (Huddy and Khatib 2007:63)

This ambiguity in definition directly impacts the empirical measurement of national identity and its different elements (Fleiß et al. 2009). On the one hand, *different items* are used to measure specific concepts, such as *constructive patriotism* (Davidov 2009; Huddy and Khatib 2007). This differential focus raises doubts about the comparability of research results that are based on a variety of indicators. On the other hand, the *same items* are used to measure different and partly contradictory concepts.

One of the most extreme examples of measuring various concepts with the same indicator is the use of the general national pride (GNP) item (see Figure 4.1). The GNP item asks respondents for a general evaluation of their national pride on a four-point scale running from "very proud" to "not proud at all." The response options "no citizen of country" and "can't choose" are given as well.

How proud are you of being [COUNTRY NATIONALITY]? (Please, check one box below.)

- | | |
|--------------------------------|--------------------------|
| Very proud | <input type="checkbox"/> |
| Somewhat proud | <input type="checkbox"/> |
| Not very proud | <input type="checkbox"/> |
| Not proud at all | <input type="checkbox"/> |
| I am not [COUNTRY NATIONALITY] | <input type="checkbox"/> |
| Can't choose | <input type="checkbox"/> |

Figure 4.1. The Implementation of the GNP Item in the 2013 ISSP

The GNP item already has served as an indicator for:

- National pride (e.g., single-item indicator [Moaddel, Tessler, and Inglehart 2008; Muñoz 2009; Tilley and Heath 2007])
- National attachment (e.g., single-item indicator [Elkins and Sides 2006])
- National identity (e.g., single-item indicator [Shayo 2009; Tiryakian 2006]; multiple items [Guinaudeau, Fuchs, and Schubert 2009; Pehrson, Vignoles, and Brown 2009])
- A construct distinct from national identity (e.g., single-item indicator [Müller-Peters 1998])
- Nationalism (e.g., single-item indicator [Solt 2011]; multiple items [Blank and Schmidt 2003; Fleiß et al. 2009; Wagner et al. 2012])
- Nationalist sentiments (e.g., single-item indicator [Han 2013])
- Patriotism (e.g., single-item indicator [Ariely 2012; Rose 1985]; multiple items [Balabanis et al. 2001; Kimmelmeier and Winter 2008; Kosterman and Feshbach 1989; Li and Brewer 2004; Sidanius et al. 1997])

Although the GNP item might be a valid measure for some concepts in some countries, it is highly improbable that this item can simultaneously measure all concepts—unless they are virtually identical—especially when used as a single-item indicator.

Given the popularity of the GNP item, this article aims to assess in an explanatory manner the suitability of the GNP item for cross-national studies of national identity and its elements. The article will present reasons for choosing the GNP item and the critical issues involved in its use. We will introduce the method of online probing (OP) and present the different concepts of national identity and its dimensions. By means of exploratory factor analysis and OP, we will demonstrate the cross-national variability of respondents' associations regarding the GNP item. We will report in detail the OP results of a web survey conducted in Germany, Great Britain, Mexico, Spain, and the U.S. The probe results revealed

respondents' various nation-related emotions, reasons, and problems when answering the GNP item. Additionally, we will assess to what degree the previously mentioned problems have a distorting impact on the answer selection at the GNP item. Finally, we will conclude with a critical discussion of the study's results.

4.1.1 The General National Pride Item and Its Usage

Several reasons exist why researchers might use this item as a single-item indicator for the different concepts of national identity. First, the implementation of this item in several cross-nation studies—such as the WVS, the ISSP, and the Eurobarometer—has provided longitudinal data (Muñoz 2009; Tilley and Heath 2007), whereas multiple indicator measures of the different constructs often are limited to specific surveys and, therefore, cover shorter time periods. Second, researchers can increase the number of countries in their study by combining different surveys, which might be interesting for studies conducting multilevel analysis or aiming at a global reach. Third, several researchers have justified the use of the GNP item as a single-item indicator (e.g., Moaddel et al. 2008; Muñoz 2009) with a study by Tilley and Heath (2007) who found for Great Britain a high correlation between the GNP item and a more sophisticated measure of national pride from the ISSP. However, two critical issues arise regarding this argumentation. On the one hand, Tilley and Heath used for this correlation ISSP items that usually serve as indicators for nationalism (Davidov 2009) or chauvinism (Coenders and Scheepers 2003), which qualifies the GNP item more as an indicator for nationalism in Great Britain. On the other hand, it remains unclear whether the correlation between the GNP item and the ISSP items in this particular case also holds for other countries or in a cross-national context. Finally, Elkins and Sides (2006) have assigned the GNP item face validity as a measure of national attachment. For them, the general scope of the GNP item makes it preferable to several nationalism and patriotism survey indicators

that are too context specific (e.g., the pride in achievements in the arts and literature or history) and which are therefore problematic as indicators for comparative studies.

However, one methodological and several substantive reasons exist that caution against the use of the GNP item. The methodological criticism rejects the very idea of a single-item indicator as a measure for complex constructs:

The underlying assumption that there is a one to one relationship between the single item and the theoretical construct and that it is measured without error (which is the implicit assumption when a scale is measured by only one item) ... is doubtful. (Ariely and Davidov 2012:273)

From a substantive point of view, two critical issues remain: missing control for effects of social desirability and missing control for respondents' lack of national identification while they identify more on a regional or global level or regard national identity as irrelevant source for their identity.

The *social desirability* issue concerns the question of how valid the GNP measure is when used for cross-national studies:

[T]here is clearly a danger that what these cross-national surveys are picking up are primarily cultural differences over how acceptable it is to express national pride, rather than varying levels of identification It appears, therefore, that using the WVS and ISSP measures of national pride to reveal cross-national differences in the strength of national identities may yield misleading results. (Miller and Ali 2014:243)

The social desirability effect potentially works in two ways. On the one hand, respondents from certain countries might feel pressured to opt for a high level of national pride because they feel that they are supposed to be proud of their country. In this study, Mexican respondents especially may show this response behavior (Klesner 2006). Moreover, this effect may be particularly problematic when the GNP item is interpreted as constructive patriotism. A previous study has revealed that Mexican respondents see the different elements of constructive patriotism as problematic (e.g., pride in democracy or the social security system)

(Meitinger, forthcoming), but rate their national pride higher when asked the GNP item (e.g., ISSP 2013 mean for “pride in democracy” 3.2 and for GNP 1.7)⁹.

On the other hand, respondents might opt for a low level of national pride, although they may feel a strong attachment to their country because it is not permissible in their country to overtly express national pride. A prime example for this scenario is Germany. In previous cross-national comparisons, Germany showed persistently low levels of national pride (Evans and Kelley 2002; Smith and Jarkko 2001) but a high level of national attachment (Miller-Idriss and Rothenberg 2012). Smith and Jarkko (2001) explained this contradiction with the “war guilt effect.” In the aftermath of WWII and the Nazi regime, a public narrative was established in West Germany that prohibited the open expression of national pride. The expression of national pride still triggers right-wing connotations to this day (Miller-Idriss 2009). However, the importance of this pride taboo and the meaning of “national pride” are shifting. Whereas the older generations still perceive “national pride” as equivalent to “rampant nationalism and racism” (Miller-Idriss and Rothenberg 2012:90), the younger generation of Germans is yearning for a more open expression of nation-related emotions, and constantly is redefining the meaning of national pride (Miller-Idriss 2009). The question remains whether the GNP item is still affected by the “pride taboo” or whether these social desirability effects are gradually disappearing in Germany.

The second substantive issue is the missing control if the respondents identify themselves sufficiently with the national level. The GNP item presupposes that the respondents indeed identify with their country, although this may not always be the case. Respondents could reject the idea of a national identity altogether, or they could mostly identify on the global or regional level. Two countries of this study—Spain and Great Britain—may especially be affected by the latter possibility. For example, Spaniards from Catalonia and the Basque Country show a lower pride in Spain than Spaniards living in other

⁹ Both items are measured on a four-point scale running from, 1 “very proud” to 4 “not proud at all.”

parts of the country. If the GNP item is used as a single indicator, it is impossible to distinguish whether these respondents identify as Spaniards and are alienated by the current Spanish nationalism or whether they identify on the regional level and the lower pride level expresses an absence of national identification (Muñoz 2009).

In a similar vein, in Great Britain, Scottish and Welsh nationalism may be a powerful source that could substitute a national with a regional identification for some respondents. In particular, Scottish respondents show lower levels of national pride than their English and Welsh counterparts (Tilley and Heath 2007).

4.1.2 National Identity and Its Elements

Despite the controversies around the theoretical conceptualization and empirical measurement of national identity and its elements, a strong research tradition exists that perceives national identity as a unifying force within societies. This perspective originated within Social Identity Theory that saw national identity as a “positive, subjectively important emotional bond with a nation” (Tajfel and Turner 1986). In a similar vein, Smith and Jarkko (2001) defined national identity as “the cohesive force that holds nations together and shapes their relationships with the family of nations” (p. 1), and for Miller and Ali (2014), it “provides the ‘cement’ or ‘glue’ that holds modern, culturally diverse, societies together and allows them to function effectively” (p. 238). However, several studies have argued that national identity is a multidimensional concept (e.g., Kosterman and Feshbach 1989) that should be divided into sub-categories, an idea that originated in Adorno and colleagues’ (1950) distinction between a love of one’s country (genuine patriotism) and an uncritical attachment to one’s country combined with a rejection of other nations (pseudo-patriotism). Several other studies have introduced dichotomizations of national identity, such as Schatz, Staub, and Lavine (1999) (blind versus constructive patriotism) and Blank, Schmidt, and Westle (2001) (nationalism versus constructive patriotism). Finally, Blank and Schmidt (2003) distinguish between

nationalism and constructive patriotism. For them, the important characteristics of nationalists are an idealization of their nation, feelings of national superiority, and an uncritical acceptance of national authorities. They reject any criticism of their nation, and their criteria for who is perceived as a member of the nation are based on descent, race, or culture; and they draw socially derogatory comparisons with groups that they do not consider part of their nation. Constructive patriots refuse an idealization of the nation and an uncritical acceptance of state authorities. They endorse its criticism, only support the nation if it is working according to humanistic and democratic principles, and cherish an advanced social system (Blank and Schmidt 2003; Davidov 2009).

Another element of national identity, of course, is national pride: “National pride involves both admiration and stake holding—the feeling that one has some kind of share in the achievement or an admirable quality” (Evans and Kelley 2002:303). However, researchers disagree about how national pride relates to patriotism and nationalism. Some researchers have equated national pride with patriotism (e.g., Rose 1985) or nationalism (e.g., Solt 2011), whereas other researchers have argued that national pride is a precondition for patriotism and nationalism:

National pride is the positive affect that the public feels towards their country, resulting from their national identity. It is both the pride or sense of esteem that a person has for one’s nation and the pride or self-esteem that a person derives from one’s national identity. ... National pride co-exists with patriotism and is a prerequisite of nationalism, but nationalism extends beyond national pride, and feeling national pride is not equivalent to being nationalistic. (Smith and Jarkko 2001:1)

The researchers who equate the two complex concepts of patriotism and nationalism with national pride may be inclined to use the GNP item as a proxy for patriotism or nationalism.

4.2 ONLINE PROBING AS A METHOD TO UNCOVER RESPONDENTS' THOUGHTS

A variety of qualitative approaches exist that can help to assess if certain items achieve cross-national comparability, such as cognitive interviewing (Willis 2005) and OP (Braun et al. 2014). A major advantage of a qualitative approach is that it can reveal the reasons for missing comparability if quantitative approaches cannot establish measurement invariance (Meitinger, forthcoming) or are inapplicable. The latter is the case when the GNP item is used as a single-item indicator because measurement invariance tests presuppose multiple indicators for one construct (Bollen 1989).

OP especially lends itself to the task at hand. OP transfers traditional cognitive interviewing techniques from the laboratory context into web surveys. During OP, web survey respondents receive verbal probes while answering a questionnaire. Probing in the context of OP means that respondents first answer a closed item, and then receive, on a second screen, follow-up questions—so-called probes—to gain further insight into the respondents' cognitive processes when they answered the initial closed item. For example, a *category-selection probe* asks the respondents why a certain answer category was chosen (Willis 2005).

The implementation of OP in web surveys allows for large sample sizes and a comparison of several countries, which permits a quantitative data analysis of qualitative insights. This implementation enables researchers to judge the prevalence of themes or error types (Meitinger and Behr, forthcoming) and to analyze specific sub-populations or response patterns. Since all respondents receive the same probe, the results are standardized (Braun et al. 2014), which can reduce some of the data harmonization issues of cross-national cognitive interviewing (Lee 2012).

4.3 RESEARCH OBJECTIVES

The main goal of this article is to assess the suitability of the GNP item as a cross-national indicator for the different elements of national identity with exploratory factor analysis and OP. The OP approach opens up a unique opportunity to reveal the various associations of respondents when they answer the GNP item. At the same time, we want to assess the prevalence of potentially problematic issues, such as effects of social desirability and the absence of national identification.

4.4 METHODS AND DATA

For the exploratory factor analysis, we used the data set from the 2013 ISSP Module on National Identity (ISSP Research Group 2015). We included in our analysis Germany ($N=1,717$), Great Britain ($N=904$), the U.S. ($N=1,274$), Mexico ($N=1,062$), and Spain ($N=1,225$), which are the countries included in our web survey.

The OP results for this study came from a web survey conducted with 2,685 participants in May 2014. The respondents from Germany, Great Britain, Mexico, Spain, and the U.S. were drawn from a non-probability online panel with quotas for age (18–30, 31–50, and 51–65), gender, and education (lower and higher).

In this web survey, we replicated questions from the ISSP module on National Identity. After the respondents answered the closed GNP item, they received, on a separate screen, a category-selection probe (see Figure 4.2) that inquired about why a certain answer category was chosen (Willis 2005). On the basis of the probe answers, we developed a coding schema. All the responses were coded by a researcher, and a randomly chosen sample of responses (20 percent) were coded a second time by a student assistant. The intercoder reliability was 93 percent, and mismatched coding was discussed by the coding team and corrected accordingly.

Please explain why you selected "not proud at all".

The question was: "How proud are you of being American?"

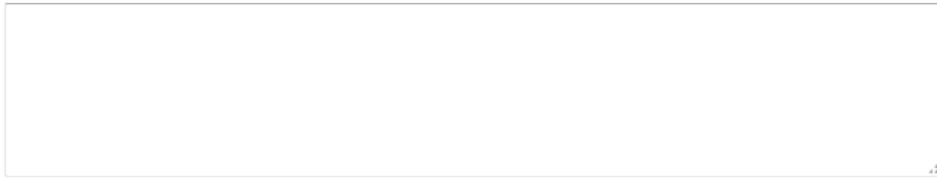


Figure 4.2. Screenshot of Category-selection Probe for the GNP Item

Since this study intends to evaluate the cross-national comparability of the GNP item from the 2013 ISSP, we considered the replication of the ISSP answer distribution of this item in our web study as a precondition for drawing comparisons between the two data sets. Table 4.1 summarizes the mean values, the standard deviations, the percentage of nonresponse (NR), and the percentage of respondents that opted for the answer category “no citizen of country” at the closed item. In both data sets, we limited our sample to respondents with the countries’ citizenships. Our web survey approximately replicated the response pattern of the 2013 ISSP, although the Spanish respondents in our web survey chose lower pride values than the Spanish ISSP respondents.

Table 4.1. Comparison of the Mean, Proportion of Nonresponse, and Proportion of the Respondents Who Chose “No Citizen of Country” of the 2013 ISSP and the Web Survey for the GNP Item

	2013 ISSP			Web Survey		
	Mean (<i>SD</i>)	NR (%)	No Citizen of Country (%)	Mean (<i>SD</i>)	NR (%)	No Citizen of Country (%)
Germany	2.1 (.7)	12.2	0	2.2 (.8)	14.1	.7
Great Britain	1.8 (.7)	2.4	1.3	1.9 (.8)	3.9	3.2
Mexico	1.7 (.8)	3.0	0	1.7 (.8)	2.1	0
Spain	1.7 (.9)	2.3	2.3	2.1 (1.0)	3.8	.7
U.S.	1.3 (.5)	1.4	.1	1.5 (.7)	2.6	.2

Note: Scale of the GNP item: four-point scale running from 1 “very proud” to 4 “not proud at all.”

4.5 RESULTS EXPLANATORY FACTOR ANALYSIS

Respondents in different countries may associate various concepts with the GNP item. For example, the GNP item may trigger patriotic associations in country A and nationalistic associations in country B. Table 4.2 is an exploratory factor analysis (EFA) of items measuring constructive patriotism (a–c), nationalism (d and e) (Davidov 2009), and the GNP item (f) for five countries using 2013 ISSP data (see Table 4.3 for the item wording) and the software package STATA 14. Since some of the items had a four-point scale (a–c, f), we used an EFA with polychoric correlations because Maximum Likelihood estimations can lead to biased parameters and standard errors when applied to ordinal scales with insufficient scale points (Schmitt 2011). If the GNP item can be a good indicator for either patriotism or nationalism, the GNP item should load in all countries on only one of the factors. However, clearly, this is not the case. In Great Britain, the GNP item could serve as an indicator for nationalism, and in Spain and the U.S., it could serve as an indicator for constructive patriotism. Even more troublesome, the GNP item does not sufficiently load on any factor in Mexico.

Table 4.2. Exploratory Factor Analysis with Items for Nationalism, Patriotism, and the GNP Item Using the 2013 ISSP

	Germany		Great Britain		Mexico		Spain		U.S.	
	Factor 1	Factor 2	Factor 1	Factor 2	Factor 1	Factor 2	Factor 1	Factor 2	Factor 1	Factor 2
a) Democracy	0.700	-0.052	0.673	0.038	0.792	-0.014	0.189	0.524	0.614	-0.030
b) Social security	0.6823	-0.086	0.618	-0.053	0.810	-0.038	-0.067	0.608	0.521	-0.089
c) Fair and equal	0.581	0.110	0.657	-0.009	0.599	0.159	0.108	0.521	0.460	0.179
d) World better place	-0.103	0.792	-0.079	0.733	-0.033	0.705	0.722	-0.035	-0.016	0.593
e) Better country	0.036	0.751	0.043	0.669	0.032	0.699	0.685	0.114	0.095	0.612
f) GNP	0.328	0.3746	0.316	0.425	0.231	0.199	0.409	0.313	0.533	0.249

Note: Polychoric correlations, principal factor, oblique rotation.

Table 4.3. Items Measuring Nationalism, Constructive Patriotism, and GNP in the ISSP 2013

Factor	Item	Question Wording
COP	a)	How proud are you of [COUNTRY] in the way democracy works?
	b)	How proud are you of [COUNTRY] in its social security system?
	c)	How proud are you of [COUNTRY] in its fair and equal treatment of all groups in society?
NAT	d)	The world would be a better place if people from other countries were more like the [COUNTRY NATIONALITY]
	e)	Generally speaking, [COUNTRY] is a better country than most other countries
GNP	f)	How proud are you of being [COUNTRY NATIONALITY]?

The EFA results already have cast doubts on the appropriateness of the GNP item as a cross-national measure for constructive patriotism or nationalism. Therefore, it is worthwhile to have a closer look what respondents think of when they answer the GNP item and to evaluate if it might serve as a cross-national indicator of any of the concepts of national identity.

4.6 INSIGHTS FROM ONLINE PROBING

The OP results give additional insights into the cross-national variability of respondents' associations regarding the GNP item.

4.6.1 The Coding Schema

Since we wanted to answer the question whether the GNP item can serve as an indicator for the different elements of national identity, we chose a twofold strategy to develop our coding schema. We based our coding schema on recent theories and empirical approaches to measuring elements of national identity. In this context, we especially followed the idea of Dekker, Malova, and Hoogendoorn (2003) that nation-related emotions should be seen as a continuum. Feeling "German" or "Spanish" and identifying with one's country serve as a precondition for liking the country, national pride, and nationalism. At the same time, respondents can have negative nation-related feelings such as shame. Since Dekker et al. (2003) did not explicitly mention the concept of patriotism in their continuum of emotions, we also based our coding schema on Blank and Schmidt's (2003) distinction between *nationalism*

and *constructive patriotism* as two sub-dimensions of national identity. We coded respondents as *constructive patriots* when they underlined the importance of democratic and humanistic values (e.g., “tolerance,” “freedom of speech”); based their answer selection on the perceived realization of democratic principles in their country (e.g., “state of the social security system,” “voting system”); or took a critical stance towards their nation, its history, or its national authorities. In contrast, we coded respondents as *nationalists* when they took an uncritical stance towards their nation, its history, and its national authorities. Respondents were also coded as nationalists when respondents perceived their nation as superior to other nations and they showed in-group favoritism; made socially derogating comparisons with groups not considered to be part of their nation; or defined their own group by criteria of descent, race, or culture. We also introduced a code for all responses that are *probably* nationalistic, but are lacking a clear nationalistic statement (*pseudo nationalist*). We also summarized expressions of national attachment, national pride, or national identification in the code *further positive national sentiments*. Additionally, we assigned the code *feelings of shame* to responses that mentioned negative nation-related feelings.

In addition to this theory driven approach, we heuristically developed further categories that capture the full range of respondents’ reasons. Respondents also mentioned the *living conditions in their country*, that they were *born in the country*, the importance of *specific values*, and *general characteristics of their country’s citizens*. They further based their evaluation on the *performance of the government or national authorities*, their *culture and traditions*, the country’s *history* or its *nature and landscape*. Several respondents were concerned about the *global reputation* of their country and its *worldwide influence*. All specific reasons that were not mentioned by at least 5 percent of the respondents in any country were summarized as *other specific reasons*. Given the already large number of different categories, we refrained from distinguishing a positive or negative evaluation of specific reasons.

The coding schema also captured different problems that appeared in the probe responses. We distinguished three main problem types. First, respondents had trouble answering this item because, for them, an *individual achievement* is a precondition to feeling proud of something. These respondents rejected the idea of being proud of a collective and more abstract achievement, which rendered impossible being proud of their country's achievements. Second, respondents pointed out that it was either unacceptable to be proud of one's country, or they associated right-wing connotations with the term pride. Thus, the effects of *social desirability* affected their judgement of pride. Third, several respondents perceived *national identity as an irrelevant* component of their identity, or they felt more connected to other levels of identification, such as the local, regional, or global. The remaining issues that the probe uncovered are summarized in the code *further problems*. All remaining substantive answers that were not mentioned by at least 5 percent of the respondents in any country were coded as *rest*.

4.6.2 Descriptive Results

The OP results indicate that respondents think about various nation-related emotions, reasons, and problems when answering the GNP item.

Nation-related emotions. This article started with the initial observation that the GNP item is used for various nation-related emotions. Therefore, we wanted to evaluate if this item is a suitable indicator for social constructs, such as *constructive patriotism* or *nationalism*. If this were the case, the majority of respondents in all countries should have opted for one particular nation-related emotion. Clearly, this is not the case. As Table 4.4 shows, respondents of all countries mentioned all types of nation-related emotions. Therefore, the general national pride item should not be used as a single-item indicator for specific constructs, such as *patriotism* or *nationalism*.

Table 4.4. Codes for the Category-selection Probe for the GNP Item in Percent

Code	Germany (N = 488)	Great Britain (N = 464)	U.S. (N = 523)	Mexico (N = 486)	Spain (N = 467)	Total (N = 2428)
Nation related emotions (%)						
Patriot	14.3	17.7	28.6	19.5	21.2	20.2
Nationalist	7.4	14.4	5.3	10.5	7.2	9.0
Pseudo nationalist	5.9	5.2	14.7	5.5	3.7	7.0
Further positive national sentiments	21.1	23.9	24.8	27.9	22.4	24.1
Feelings of shame/absence of pride	6.2	6.3	3.4	2.9	14.6	6.6
Specific reasons (%)						
Living conditions	12.3	10.1	8.3	21.0	19.8	14.5
Born in the country	9.4	7.3	9.2	26.4	16.1	14.0
Specific values	12.1	18.5	26.5	11.9	11.9	16.0
General characteristics of country's citizens	4.7	8.4	4.7	19.5	15.6	10.8
Performance of the government or national authorities	8.6	11.9	9.8	19.9	31.9	16.6
Culture and traditions	4.5	7.5	1.3	17.0	9.7	8.2
History	14.3	8.8	1.7	8.6	6.6	8.1
Nature and landscape	2.1	2.2	.4	16.8	5.4	5.6
Global reputation	7.4	6.0	1.5	4.2	5.6	4.9
Worldwide influence	6.6	8.2	5.1	.4	.6	4.1
Other specific reasons	2.5	9.3	2.4	1.2	7.2	4.4
Problems with question (%)						
Missing individual achievement	10.7	1.1	.6	.2	2.1	2.9
Effects of social desirability	4.3	1.9	0	0	.4	1.3
Missing relevance of national identity	10.3	11.6	1.1	1.0	8.6	6.4
Further problems	5.1	1.1	.6	.4	1.0	1.7
Rest (%)						
Rest	10.8	10.9	11.7	4.7	5.2	8.6

Spanish respondents more often expressed shame than respondents from other countries. Most of these respondents were dissatisfied with a political class they perceived as corrupt:

As I feel slightly more ashamed for the political class every day because nobody has the modesty to resign or to admit his or her mistakes. Their priority is to steal as much as possible during their legislative period. (Spaniard, "not very proud")

I cannot be proud to belong to a country that has an unbelievably corrupt political class. (Spaniard, "not very proud")

Although the Spanish media increased the perception of corruption by their intensive coverage of the latest corruption scandals (Marek 2015), democracy in Spain is still more stable than in Mexico (Freedom House 2015). Interestingly, fewer Mexican respondents expressed feelings of shame towards their nation than respondents from any other country. This result may be an indication that for Mexican respondents, it is socially undesirable to express shame regarding their country.

Hidden patriots and nationalists. A second issue regarding nation-related emotions exists. A common assumption of studies using the GNP item as a single-item indicator for nationalism or patriotism is that they define nationalists or patriots as all the respondents who opted for the answer values “very proud” or “somewhat proud.” However, the probing results revealed that this is a questionable assumption. In total, 493 respondents were coded as patriots and 219 respondents were coded as nationalists (see Table 4.5). Several of these respondents chose the answer values “not very proud,” “not proud at all,” or “can’t choose” at the closed GNP item. Following the usual approach, these respondents would not be included in the calculation of patriots and nationalists. Therefore, we call these respondents *hidden patriots* and *hidden nationalists* (respondents coded as patriots or nationalists in the probe responses but who chose an unexpected answer value at the closed GNP item).

Hidden patriots and hidden nationalists have different reasons for choosing the seemingly contradictory answer values at the closed GNP item. The majority of hidden patriots complain about the current government, and this is especially the case in Mexico: “I am proud to be Mexican. I am not proud of the government and what is happening in Mexico. They make our country look bad” (Mexican, “not very proud”).

A few German respondents classified as hidden patriots opted for the “don’t know” answer option because they were torn between being proud of the state of democracy and their critical perception of Germany’s history: “I am proud to live in a democratic country. ...

I am not proud of the history with both wars” (German, “can’t choose”). By contrast, hidden nationalists are driven by in-group favoritism, mostly criticizing the perceived priority treatment of foreigners in their country.

In our study, although just 6 percent of all constructive patriots and 12 percent of all nationalists fell in the category of hidden patriot or hidden nationalist, a considerable variation exists across countries. The issue of hidden patriots is more prevalent in Germany, Mexico, and Spain, whereas the probe revealed hidden nationalists mostly in Great Britain, and in particular, Germany. For example, a reliance on the closed GNP item would underestimate the percentage of German nationalists by one third. This number might even be higher because we did not account for pseudo nationalists in this calculation.

Table 4.5. Percentage of Hidden Patriots and Hidden Nationalists (n)

	Germany	Great Britain	U.S.	Mexico	Spain	Total
Respondents coded as patriots (<i>N</i>)	70	82	134	104	103	493
“Hidden patriots” (%) (<i>n</i>)	9 (6)	4 (3)	0	10 (11)	11 (11)	6 (31)
Respondents coded as nationalist (<i>N</i>)	36	67	25	56	35	219
“Hidden nationalists” (%) (<i>n</i>)	33 (12)	16 (11)	12 (3)	2 (1)	0	12 (27)

Specific reasons. Many respondents mentioned specific reasons for their nation-related emotions or to justify their pride selection without referring to a nation-related emotion. Although the respondents from all countries provided various specific reasons, some seemed to be country specific, and the number of relevant factors also differed across countries. For example, the U.S. respondents were concerned mostly with specific values (e.g., *freedom, liberty*) but none of the other reasons was mentioned by more than 10 percent of U.S. respondents. American respondents more directly expressed nation-related emotions but infrequently gave specific reasons to explain their pride evaluation. The importance of specific values also was mentioned by British (e.g., *tolerance, multiculturalism*) and German respondents (e.g., *freedom of speech, (in)equality, social security*), but they also thought about the current living conditions in their country. The performance of the government and national authorities also was a relevant factor for the British respondents, whereas the German

respondents more often explained their modest level of pride with reference to Germany's history. Since these German respondents often refer to WW2 and the Nazi regime in their probe responses, this is a strong indication that the "war guilt" effect (Smith and Jarkko 2001) still influences their pride evaluation.

In contrast, Spanish and Mexican respondents thought about a wider variety of specific reasons. Similar to other countries, the Spanish speaking respondents referred to living conditions and values (e.g., Mexico: *freedom of choice, freedom of speech*; Spain: *(in)equality, family*). A more detailed analysis of the probe responses revealed that they also were discontented with the current government and national authorities, especially the Spaniards. All Spanish probe responses that were assigned the code "performance of the government or national authorities" took a negative stance, which showed that the Spaniards were affected by a high level of political disenchantment (31.9 percent of Spanish respondents). They complained about the prevalence of corruption, the shortcomings of the judiciary system, and cutbacks in the social sector. This result is in line with previous research indicating that the economic crisis in Spain has led to an increased level of political distrust due to a negative perception of the political responsiveness of representative institutions, and an increasing perception of political corruption (Torcal 2014) that has been fostered by extensive media coverage (Marek 2015). These results question the assumption that the GNP item is superior to more specific pride items because it is less affected by context effects (Elkins and Sides 2006). The pride level of the Spanish respondents with respect to the GNP item is clearly affected by the economic crisis, which is a context effect. Furthermore, being born in the country was a response provided by about one quarter of the Mexican respondents and about one sixth of the Spanish respondents as a reason for their pride evaluation. Additionally, in both countries, the respondents pointed to the typical character traits of their fellow citizens. However, the Mexican respondents mentioned two reasons that appeared less frequently in the responses from those from other countries: First, the Mexican respondents

underlined the importance of culture and tradition (17.0 percent of Mexican respondents), in particular traditional Mexican cuisine. Second, the probe also revealed that the landscape, nature, and climate were central features of Mexican national pride (16.8 percent of Mexican respondents).

Problems. More troublesome than the previous issues is the percentage of respondents indicating one of the following problem types. Since about one quarter of the German respondents mentioned at least one problem type, the GNP item seems to be particularly troublesome in the German context.

Missing individual achievement

In particular, the problem code “missing individual achievement” seems to be a country-specific issue restricted largely to the German respondents. More than one tenth of all German respondents denied the possibility of being proud of their country. For them, being proud presupposes an individual achievement or contribution that is only possible on the personal, but not the national, level.

I cannot be proud of something that I did not work hard for. It is a coincidence that I am German. (German, “not very proud”)

I cannot be proud of something where I haven’t had any influence at all. (German, “not proud at all”)

This has nothing to do with me, though. After all, I have not contributed anything. (German, “can’t choose”)

This position follows the lead of the former German President Johannes Rau. In 2001, the social acceptance of national pride was fiercely discussed by the political elite in Germany (“Nationalstolzdebatte”). Rau rejected the notion of a German national pride:

You cannot be proud of something that you did not achieve yourself but you should be glad or thankful to be German. However, you cannot be proud of it, to the best of my belief. You are proud of something that you accomplished yourself. (as cited in Häusler 2002:144)

Apparently, this position still strongly influences the German respondents' perception of national pride.

Effects of social desirability

Closely related to the previous issue is the appearance of the effects of social desirability, which were mentioned by the British and German respondents. A content analysis of the probe responses that were assigned the code "social desirability" revealed that respondents from both countries associated with the term pride right-wing connotations:

It's not that I don't feel a sense of connectedness to my country or its people, I just see "Britishness" or the "pride" often as euphemisms for jingoism or racism. I like where I live and my culture, but I don't like the way that "pride" in our country has become a way for right wing groups to exclude others. (British, "not very proud")

Whereas, in this context, British respondents referred only to the right-wing association with *pride*, several German respondents perceived the expression of national pride as something forbidden: "We still should not be too proud of our country" (German, "somewhat proud") and "Being proud with this history? That is very easily misunderstood" (German, "not very proud").

These remarks reveal the persisting effect of the "pride taboo" (Miller-Idriss 2009). What Smith and Jarkko (2001) described as the "war guilt effect" still prevents Germans from freely expressing national pride and thus affects their answer selection at the GNP item. As a German respondent pointed out:

The term pride does not fit any of these questions. ... Maybe one should substitute the term with to be glad or to be happy, then my answers would not all have been negative. The results of the survey will certainly be distorted through the term pride. (German, "not proud at all")

Missing relevance of national identity as source for pride

A third issue related to the GNP item is that some respondents rejected the national level as a primary source for pride and their identity. Although about one tenth of German, British, and Spanish respondents mentioned this problem, the reasons differ across countries. The German

respondents often maintained that the concept of national pride is irrelevant for their identity: “As for me it does not matter which nationality someone has. It should have been possible to choose ‘I do not care’ as an answer value” (German, “can’t choose”).

A few German respondents also perceived themselves more as world citizens and saw the global level as a primary source of identification: “I do have my roots in Germany, but I am proud of being a world citizen” (German, “can’t choose”).

British and Spanish respondents also mentioned the irrelevance of national pride and identification as world citizens in their responses. However, the majority of British and Spanish respondents coded in this category emphasized that they identified more with a regional, rather than national, level. British respondents characterized themselves more as English or Welsh, rather than British. However, the strongest regional identification came from the respondents who identified themselves as Scottish: “I am Scottish and resent the fact that Britain has become synonymous with England in the eyes of both the English and, by dint of the BBC, the wider world as well” (British, “not proud at all”).

In a similar vein, several Spaniards preferred the regional level of Catalonia as a source of identification, rather than the national level: “I am proud of being Catalanian, and I deeply regret to have been born with Spanish citizenship. I hope this will change soon” (Spanish, “not proud at all”).

Given the strong regional identities of the Scots and Catalonians, it might not come as a surprise that several respondents rejected the national level as a source of identity. This result also is in line with previous research on national pride in Great Britain and Spain (Munoz 2009; Tilley and Heath 2007). Therefore, when using the GNP item for data analysis, it is important to be aware of these issues and, depending on the research question, it is necessary to control for these respondents. A feasible solution is to include questions that inquire about the importance of the different levels of identification (e.g., in the ISSP) or to use the Moreno scale (Moreno 1988).

In total, our OP revealed several problematic issues. However, the more pressing question is to what degree these issues have a distorting impact on the answer selection of the GNP item.

4.6.3 Which Factors Influence the Answer Selection of the GNP Item?

To answer this question, we conducted an additional regression analysis of the GNP item on the probe codes with STATA 14 (see Table 4.6). Since the dependent variable (GNP item) is measured with a four-point scale, it is necessary to account for ordinality. Instead of an ordinary least square (OLS) regression, we used an ordered logit (OLOGIT) regression. Applying the proportional odds model, we assumed that the scale of the GNP item represents a rough measure of an underlying latent, continuous scale. Since we excluded all the respondents who gave a nonresponse or “no citizen of country” response for the GNP item or a probe nonresponse, the sample size was reduced to 2,318 respondents. We also reverse coded the GNP item to facilitate interpretation. Now, the higher the answer value is, the higher is the pride level. Since our primary goal was to evaluate the distorting impact of the three problem types, we started by including dummy variables for *missing achievement*, *social desirability*, and *irrelevance for identity* (Model 1). We further controlled for age, gender, education, and country (Model 2) and nation-related emotions (Model 3). In our final model, we also added specific reasons for the pride evaluation (Model 4). *McKelvey & Zavoina's R²¹⁰* increased with every model, which indicated that the model was improving. Since we were calculating an OLOGIT regression, the interpretation of the regression was not as straightforward as an ordinary OLS regression, but the effect direction and the significance level could be interpreted. Most of the variables have a significant or highly significant effect in Model 4 (except age, education, and the characteristics of citizens). Additionally, the three

¹⁰ *McKelvey & Zavoina's R²* provides a close approximation of the R² that would be obtained when fitting the linear regression model of the underlying latent, continuous variable (Long and Freese 2014).

problem types in all four models have a highly significant negative effect on the *pride* evaluation.

Table 4.6. Ordered Logit Regression Analysis of the GNP Item on Problem Types, Background Variables, Nation-Related Emotions, and Specific Reasons (Standard Error in Parentheses)

Variable	Model 1 Problem types	Model 2 + Background variables	Model 3 + Nation related emotions	Model 4 + Specific reasons
Problem types (reference: not mentioned)				
Problem: missing achievement	-2.488 (.240)***	-2.085 (.249)***	-1.758 (.260)***	-2.297 (.274)***
Problem: social desirability	-1.558 (.366)***	-1.194 (.376)***	-.985 (.391)**	-1.172 (.403)***
Problem: irrelevance for identity	-1.956 (.175)***	-1.782 (.178)***	-1.331 (.189)***	-1.688 (.206)***
Background variables:				
Age		.017 (.003)***	.016 (.003)***	.016 (.003)***
Men (reference: women)		-.022 (.079)	.037 (.082)	-.004 (.084)
High education (reference: lower)		-.171 (.080)**	-.096 (.083)	-.083 (.085)
Country (reference: the U.S.)				
Country: Germany		-1.476 (.133)***	-1.513 (.140)***	-1.546 (.146)***
Country: Great Britain		-.851 (.130)***	-.885 (.138)***	-.799 (.141)***
Country: Mexico		-.578 (.126)***	-.591 (.134)***	-.731 (.147)***
Country: Spain		-1.386 (.131)***	-1.187 (.136)***	-1.011 (.145)***
Nation related emotion (reference: not mentioned)				
Patriot			1.384 (.125)***	2.248 (.153)***
Nationalist			1.918 (.169)***	1.877 (.176)***
Pseudo-nationalist			2.128 (.190)***	1.913 (.194)***
Positive nation related emotions			1.759 (.120)***	1.467 (.123)***
Shame/indifference			-1.806 (.177)***	-1.562 (.183)***
Specific reasons (reference: not mentioned)				
Reason: living conditions				-.285 (.122)**
Reason: born/roots here				.638 (.131)**
Reason: values				-.474 (.137)***
Reason: characteristics citizens				-.072 (.140)
Reason: national authorities				-1.627 (.133)***
Reason: culture/tradition				.594 (.172)***
Reason: history				-.312 (.164)*
Reason: nature/landscape				.698 (.210)***
Reason: outside perception				-.365 (.190)*
Reason: power/influence				-.631 (.210)***
Reason: other specific reasons				-.391 (.213)*
_cut 1	-3.139 (.097)	-3.482 (.182)	-3.134 (.204)	-3.710 (.216)
_cut 2	-1.609 (.058)	-1.913 (.164)	-1.215 (.183)	-1.611 (.191)
_cut 3	.384 (.044)	.221 (.159)	1.431 (.186)	1.289 (.191)
McKelvey & Zavoina's R ²	.108	.191	.396	.474
LR chi ² (df)	256.45 (3)	467.23 (10)	1053.96 (15)	1309.56 (26)

Notes: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$; $N = 2,318$.

To better assess the impact of the three problem types, we estimated the predicted probabilities for the countries in which each issue was most prevalent. The predicted probabilities provide the probability of choosing a certain answer value while holding the other variables constant at their mean values. As can be seen in Table 4.7, a German respondent's mention of any of the three problem types increased the probability for choosing the answer value "not proud at all" and decreased the probability for choosing the answer value "very proud." In Great Britain and Spain, respondents referring to the problem type "irrelevance for identity" also had a higher probability to opt for low pride values than respondents who did not mention this issue.

Table 4.7. Predicted Probabilities of GNP Values in Different Countries by Problem Type "Not Mentioned" versus "Mentioned"

Germany	Individual Achievement		Social Desirability	
	Not Mentioned	Mentioned	Not Mentioned	Mentioned
Not at all	.05	.36	.06	.16
Not very	.26	.46	.27	.45
Somewhat	.58	.17	.57	.36
Very	.11	.01	.10	.03

Irrelevance for Identity	Germany		Great Britain		Spain	
	Not Mentioned	Mentioned	Not Mentioned	Mentioned	Not Mentioned	Mentioned
Not at all	.05	.23	.03	.14	.03	.16
Not very	.26	.48	.17	.43	.19	.45
Somewhat	.58	.27	.62	.39	.62	.36
Very	.11	.02	.19	.04	.16	.03

4.7 DISCUSSION

This study set out to evaluate the suitability of the GNP item as a measurement for national identity and its different elements in cross-national studies. Our results clearly question the assumption that the GNP item can serve as a proxy for *nationalism* or *patriotism* in cross-national studies, since our respondents associated various concepts with this item. In addition, the GNP item may underestimate the proportion of nationalists and patriots, in particular in

Germany. More troublesome, we uncovered several problem types that distorted the answer selection of our German, British, and Spanish respondents.

These results caution against the use of the GNP item in a cross-national context, and instead strongly call for a measurement of the different elements of national identity with multiple indicators based on a strong theoretical foundation and definition. For example, Davidov (2009) developed multiple indicator measures for *constructive patriotism* and *nationalism* that showed metric invariance for 34 countries. The factor scores of these measures could, for example, be used in a cross-national regression analysis. Additionally, Hjerm and Schnabel (2010) developed a multiple indicator measure for *national sentiments* that achieved metric invariance for European countries. In contrast to single-item indicators, these multiple-item measures allow for an evaluation of cross-national comparability (Bollen 1989). As our results revealed, the use of single-item indicators, such as the GNP item, is problematic in cross-national studies.

Additionally, the probe answer revealed that respondents from different countries justify their answer selection with various reasons. Especially, the Mexican respondents provided several distinctive reasons that were less frequently mentioned by the respondents from other countries. Although these findings are not necessarily an issue for cross-national comparability in itself, they reflect the fact that the elements constituting national pride are manifold and can vary across countries. However, these results have uncovered a more general problem associated with the measurement of national pride in cross-national surveys, such as the ISSP Module on National Identity. This module also asks respondents about their pride in 10 specific domains, such as the country's history, its social security system, and the state of its economy. Although *tradition* and *nature* represent relevant elements of Mexican national pride, the ISSP module does not contain items about these specific domains of national pride. Given that the modules for these large-scale cross-national surveys are usually developed in the U.S. or Europe, a risk exists that important features of national pride will be

missed in countries outside the Western hemisphere. Skjåk (2010) already has observed this phenomenon:

Regardless of how thorough and academically rigorous the development of the methodologies, theories, concepts, and instruments are, there is always a danger that the focus on relatively homogenous countries in the beginning years of the ISSP could result in cultural constraints and ethnocentric bias when the studies expand to new cultural areas. (P. 504)

Once again, our OP study has revealed just how challenging the cross-national measurement of national identity and its elements can be. To grasp its full complexity, this challenge should be met by a combined approach of quantitative methods and qualitative approaches, such as cognitive interviewing or OP.

References

- Adorno, Theodor W., Else Frenkel-Brunswik, Daniel J. Levinson, and R. Nevitt Sanford. 1950. *The Authoritarian Personality*. Oxford: Harpers.
- Ariely, Gal. 2012. "Globalization, Immigration and National Identity: How the Level of Globalization Affects the Relations between Nationalism, Constructive Patriotism and Attitudes Toward Immigrants?" *Group Processes & Intergroup Relations* 15(4):539–57.
- Ariely, Gal and Eldad Davidov. 2012. "Assessment of Measurement Equivalence with Cross-national and Longitudinal Surveys in Political Science." *European Political Science* 11(3):363–77.
- Balabanis, George, Adamantios Diamantopoulos, Rene Dentiste Mueller, and T. C. Melewar. 2001. "The Impact of Nationalism, Patriotism and Internationalism on Consumer Ethnocentric Tendencies." *Journal of International Business Studies* 32(1):157–75.
- Behr, Dorothee, Michael Braun, Lars Kaczmirek, and Wolfgang Bandilla. 2014. "Item Comparability in Cross-national Surveys: Results from Asking Probing Questions in Cross-national Web Surveys about Attitudes towards Civil Disobedience." *Quality & Quantity* 48(1):127–48.
- Blank, Thomas and Peter Schmidt. 2003. "National Identity in a United Germany: Nationalism or Patriotism? An Empirical Test with Representative Data." *Political Psychology* 24:289–311.
- Blank, Thomas, Peter Schmidt, and Bettina Westle. 2001. "Patriotism—A Contradiction, a Possibility or an Empirical Reality." *European Consortium for Political Research Joint Sessions Institute of Political Studies, Grenoble, 6–11 April 2001*. Colchester, UK: ECPR.
- Bollen, Kenneth A. 1989. *Structural Equations with Latent Variables*. New York, NY: John Wiley & Sons.

- Braun, Michael, Dorothée Behr, Lars Kaczmirek, and Wolfgang Bandilla. 2014. "Evaluating Cross-national Item Equivalence with Probing Questions in Web Surveys." Pp. 184-200 in *Improving Survey Methods: Lessons from Recent Research*, edited by U. Engel, B. Jann, P. Lynn, A. Scherpenzeel, and P. Sturgis. New York: Routledge.
- Coenders, Marcel and Peer Scheepers. 2003. "The Effect of Education on Nationalism and Ethnic Exclusionism: An International Comparison." *Political Psychology* 24(2): 313-43.
- Davidov, Eldad. 2009. "Measurement Equivalence of Nationalism and Constructive Patriotism in the ISSP: 34 Countries in a Comparative Perspective." *Political Analysis* 17(1):64-82.
- De Figueiredo, Rui J.P. and Zachary Elkins. 2003. "Are Patriots Bigots? An Inquiry into the Vices of In-group Pride." *American Journal of Political Science* 47(1):171-88.
- Dekker, Henk, Darina Malova, and Sander Hoogendoorn. 2003. "Nationalism and Its Explanations." *Political Psychology* 24(2):345-76.
- Elkins, Zachary and John Sides. 2006. *In Search of the Unified Nation-state: National Attachment among Distinctive Citizens*. Irvine, CA: Center for the Study of Democracy. Retrieved October 13, 2015 (<https://escholarship.org/uc/item/20f203bx>).
- Evans, Maria D. and Jonathan Kelley. 2002. "National Pride in the Developed World: Survey Data from 24 Nations." *International Journal of Public Opinion Research* 14(3):303-38.
- Fleiß, Jürgen, Franz Höllinger, and Helmut Kuzmics. 2009. "Nationalstolz zwischen Patriotismus und Nationalismus?" *Berliner Journal für Soziologie* 19(3):409-34.
- Freedom House. 2015. *Freedom in the World 2014: The Annual Survey of Political Rights and Civil Liberties*. New York, NY: Rowman & Littlefield.

- Guinaudeau, Isabelle, Dieter Fuchs, and Sophia Schubert. 2009. "National Identity, European Identity and Euroscepticism." Pp. 91–112 in *Euroscepticism: Images of Europe among Mass Publics and Political Elites*, edited by F. Dieter, M. B. Raul, and R. Antoine. Farmington Hills, MI: Barbara Budrich Publishers.
- Han, Kyung Joon. 2013. "Income Inequality, International Migration, and National Pride: A Test of Social Identification Theory." *International Journal of Public Opinion Research* 25(4):502–21.
- Häusler, Alexander. 2002. "Die 'Nationalstolz'-Debatte als Markstein einer Rechtsentwicklung der Bürgerlichen Mitte." Pp. 123–46 in *Themen der Rechten—Themen der Mitte: Zuwanderung, Demografischer Wandel und Nationalbewusstsein*, edited by C. Butterwegge, J. Cremer, A. Häusler, G. Hentges, T. Pfeiffer, C. Reißlandt, and S. Salzborn. Opladen: Leske + Budrich.
- Hjerm, Mikael and Annette Schnabel. 2010. "Mobilizing Nationalist Sentiments: Which Factors Affect Nationalist Sentiments in Europe?" *Social Science Research* 39(4): 527–39.
- Huddy, Leonie and Nadia Khatib. 2007. "American Patriotism, National Identity, and Political Involvement." *American Journal of Political Science* 51(1):63–77.
- ISSP Research Group. 2015. *International Social Survey Programme: National Identity III—ISSP 2013*. ZA5950 Data file Version 2.0.0. Cologne: GESIS Data Archive. doi: 10.4232/1.12312.
- Kimmelmeier, Markus and David G. Winter. 2008. "Sowing Patriotism, but Reaping Nationalism? Consequences of Exposure to the American Flag." *Political Psychology*, 29(6):859–79.
- Klesner, Joseph L. 2006. "Economic Integration and National Identity in Mexico." *Nationalism and Ethnic Politics* 12(3–4):481–507.

- Kosterman, Rick and Seymour Feshbach. 1989. "Toward a Measure of Patriotic and Nationalistic Attitudes." *Political Psychology* 10(2):257–74.
- Latcheva, Rossalina. 2011. "Cognitive Interviewing and Factor-analytic Techniques: A Mixed Method Approach to Validity of Survey Items Measuring National Identity." *Quality & Quantity* 45(6):1175–99.
- Lee, Jihyun. 2012. "Conducting Cognitive Interviews in Cross-national Settings." *Assessment* 21:227–40.
- Li, Qiong and Marilyn B. Brewer. 2004. "What Does It Mean to Be an American? Patriotism, Nationalism, and American Identity after 9/11." *Political Psychology* 25(5):727–39.
- Long, J. Scott and Jeremy Freese. 2014. *Regression Models for Categorical Dependent Variables Using Stata*. 3rd ed. College Station, TX: Stata Press.
- Malešević, Siniša. 2011. "The Chimera of National Identity." *Nations and Nationalism* 17(2):272–90.
- Marek, Jennifer. 2015. "The Effects of Scandals on Perceived Corruption in Spain." *Global Journal on Humanities and Social Sciences* 2(1):1–8.
- Meitinger, Katharina (forthcoming): "Necessary but Insufficient: Why Measurement Invariance Tests Need Online Probing as a Complementary Tool."
- Meitinger, Katharina and Dorothee Behr (forthcoming). "Comparing Cognitive Interviewing and Online Probing: Do They Find Similar Results?" *Field Methods*.
- Miller, David and Sundas Ali. 2014. "Testing the National Identity Argument." *European Political Science Review* 6(2):237–59.
- Miller-Idriss, Cynthia. 2009. *Blood and Culture: Youth, Right-wing Extremism, and National Belonging in Contemporary Germany*. Durham: Duke University Press.
- Miller-Idriss, Cynthia and Bess Rothenberg. 2012. "Ambivalence, Pride and Shame: Conceptualizations of German Nationhood." *Nations and Nationalism* 18(1):132–55.

- Moaddel, Mansoor, Mark Tessler, and Ronald Inglehart. 2008. "Foreign Occupation and National Pride. The Case of Iraq." *Public Opinion Quarterly* 72(4):677–705.
- Moreno, Luis. 1988. "Scotland and Catalonia: The Path to Home Rule." *The Scottish Government Yearbook*:166–82.
- Müller-Peters, Anke. 1998. "The Significance of National Pride and National Identity to the Attitude toward the Single European Currency: A Europe-wide Comparison." *Journal of Economic Psychology* 19(6):701–19.
- Muñoz, Jordi. 2009. "From National-Catholicism to Democratic Patriotism? Democratization and Reconstruction of National Pride: The Case of Spain (1981–2000)." *Ethnic and Racial Studies* 32(4):616–39.
- Pehrson, Samuel, Vivian L. Vignoles, and Rupert Brown. 2009. "National Identification and Anti-immigrant Prejudice: Individual and Contextual Effects of National Definitions." *Social Psychology Quarterly* 72(1):24–38.
- Reeskens, Tim and Matthew Wright. 2013. "Nationalism and the Cohesive Society: A Multilevel Analysis of the Interplay among Diversity, National Identity, and Social Capital across 27 European Societies." *Comparative Political Studies* 46(2):153–81.
- Rose, Richard. 1985. "National Pride in Cross-national Perspective." *International Social Science Journal* 37(1):85–96.
- Shayo, Moses. 2009. "A Model of Social Identity with an Application to Political Economy: Nation, Class, and Redistribution." *American Political Science Review* 103(2):147–74.
- Schatz, Robert T., Ervin Staub, and Howard Lavine. 1999. "On the Varieties of National Attachment: Blind versus Constructive Patriotism." *Political Psychology* 20(1): 151–74.
- Schmitt, Thomas A. 2011. "Current Methodological Considerations in Exploratory and Confirmatory Factor Analysis." *Journal of Psychoeducational Assessment* 29(4): 304–21.

- Sidanius, Jim, Seymour Feshbach, Shana Levin, and Felicia Pratto. 1997. "The Interface between Ethnic and National Attachment: Ethnic Pluralism or Ethnic Dominance?" *Public Opinion Quarterly* 61(1):102–33.
- Skjåk, Knut Kalgraff. 2010. "The International Social Survey Programme: Annual Cross-national Social Surveys Since 1985." Pp. 497–506 in *Survey Methods in Multinational, Multiregional, and Multicultural Contexts*, edited by J. A. Harkness, M. Braun, B. Edwards, T. P. Johnson, L. Lyberg, P. Ph. Mohler, B. Pennell, and T. W. Smith. Hoboken, NJ: Wiley-Blackwell.
- Smith, Anthony D. 1991. *National Identity*. London: Penguin Books.
- Smith, Tom W. and Lars Jarkko. 2001. *National Pride in Cross-national Perspective*. National Opinion Research Center. Chicago: National Opinion Research Center. Retrieved August 15, 2015 (www.issp.org/documents/natpride.doc).
- Smith, Tom W. and Benjamin Schapiro. 2015. *A Compilation of Documents Used to Develop the National Identity III Questionnaire for the International Social Survey Programme in 2014*. NORC. Chicago: U.S. Retrieved October 10, 2015 (<http://www.gesis.org/issp/modules/issp-modules-by-topic/national-identity/2013/>).
- Solt, Frederick. 2011. "Diversionary Nationalism: Economic Inequality and the Formation of National Pride." *The Journal of Politics* 73(3):821–30.
- Tajfel, Henri and John C. Turner. 1986. "The Social Identity Theory of Intergroup Behavior." Pp. 7–24 in *Psychology of Intergroup Relations*, edited by W. G. Austin and S. Worchel, Chicago: Nelson-Hall.
- Tilley, James and Anthony Heath. 2007. "The Decline of British National Pride." *The British Journal of Sociology* 58(4):661–78.
- Tiryakian, Edward A. 2006. "Coping with Collective Stigma: The Case of Germany." Pp. 359–98 in *Identity, Morality, and Threat: Studies in Violent Conflict*, edited by D. Rothbart and K. Korostelina. Lanham: Lexington/Rowman & Littlefield.

- Torcal, Mariano. 2014. "The Decline of Political Trust in Spain and Portugal: Economic Performance or Political Responsiveness." *American Behavioral Scientist* 58(12):1542–67.
- Wagner, Ulrich, Julia C. Becker, Oliver Christ, Thomas F. Pettigrew, and Peter Schmidt. 2012. "A Longitudinal Test of the Relation between German Nationalism, Patriotism, and Outgroup Derogation" *European Sociological Review* 28(3):319–32.
- Willis, Gordon B. 2005. *Cognitive Interviewing: A Tool for Improving Questionnaire Design*. Thousand Oaks: Sage Publications.

5. General Conclusion

This dissertation project set out to explore the potential of the method of online probing in comparison to the methods of cognitive interviewing and MGCFA, and to evaluate the potential of online probing for assessing single indicator measures. At the same time, the study aimed to assess whether online probing can fulfill its goals of understanding cognitive processes and, in an international context, shedding light on the equivalence of survey items. The substantive applications used for these explorations were the item battery on specific national pride, the constructs of nationalism and constructive patriotism, and the general national pride item.

5.1 ARTICLE 1: RESULTS, LIMITATIONS, AND FUTURE RESEARCH

Since online probing and cognitive interviewing use the same technique (probing), they are very similar methods. At the same time, they have different mode settings, diverging sample sizes, varying levels of interactivity, and research goals. The first article (Chapter 2) in this dissertation chose the item battery on specific national pride as a substantive application and answered the following research questions:

1. Are there indications that response quality differs between CI and OP?
2. Do cognitive interviewing and online probing methods produce similar results? How do sample size and the divergent levels of interactivity affect the performance of these two qualitative methods?

This article compared online probing and cognitive interviewing from an error and theme perspective, since these standpoints represent the respective research goal of both methods. The error perspective evaluated the methods' potential for detecting problems at the

different items and the theme perspective compared both methods in regard to the content and variety of the associations that are mentioned by respondents of these methods.

5.1.1 Results

The results of this study found that cognitive interviewing achieved a higher response quality than online probing in the probe answers. During cognitive interviewing, respondents opted less often for an item nonresponse and gave longer responses. However, the results had a local bias, since the cognitive interviewing respondents came from the metropolitan area of Mannheim. In contrast, the implementation in the web survey allowed for a nationwide sampling, which prevented a local bias in online probing.

Both methods also had an extensive overlap of results with respect to revealed error types and uncovered themes. The respondents of the cognitive interviewing mentioned more error types and were slightly more productive in theme perspective than the respondents of online probing. However, the large sample size of online probing compensated for the lower productivity of its respondents and enabled a judgement about the prevalence of error types. At the same time, the presence of an interviewer during the cognitive interviewing increased interactivity and allowed for the spontaneous remarks of respondents, which clearly increased the performance of this method with respect to error detection but not theme detection.

In addition to answering the two main research questions, this article also gave some practical guidance in regard to the expected work load and several key characteristics of both methods, which will help researchers to decide which method may be more appropriate to use in a particular research situation. Given the complimentary strength of both methods, this article also discussed the potentials of combining online probing and cognitive interviewing in a single study, and therefore, encouraging researchers to conduct qualitative mixed-methods studies.

5.1.2 Limitations

The overall goal of this article was to compare the probing in cognitive interviewing and online probing when both are conducted “as usual.” Two clear limitations follow from this research goal. First, the respondents of the two methods received different briefings, which may potentially explain the higher productivity of the respondents of the cognitive interviewing in finding errors. Second, the sample size of the methods differed (cognitive interviewing: 20 respondents; online probing: 532 respondents split into five groups). Although this article served its purpose to assess the “usual” approach of both methods, future research should test online probing with different briefings with respect to error detection, and it also should test the potential of both methods when equal sample sizes are used.

5.1.3 Future research

This article also revealed two relevant issues for future research: interactivity and the reduction of workload in the coding process of online probing. First, the method comparison showed the importance of *interactivity* in cognitive interviewing and, in particular, the motivating effect of the cognitive interviewers. In contrast, online probing was affected by item nonresponse at varying levels and mismatching answers which are well-known issues associated with online probing (Braun et al. 2015). In cooperation with Lars Kaczmirek and Dorothée Behr, I developed a tool that can automatically detect instances of probe nonresponses in web surveys. If a probe nonresponse is detected, the survey program repeats the probe along with a motivational sentence that is adapted to the type of probe nonresponse that was given (e.g., the tool can distinguish between nonresponses where respondents gave don't know answers, when they refused to answer or gave non-sense answers). This tool enabled the survey software to detect cases of probe nonresponse and to trigger a conditional probe in the web study of the first article. Although the tool could not improve the error detection of OP in this study, it shows greater potential in the reduction of probe nonresponse

in theme detection. When tested with 30 different items, the tool converted, on average, 56 percent of probe nonresponses into substantive responses, with some items reaching a conversion rate of up to 74 percent (Kaczmirek, Meitinger, and Behr 2015). However, the automatic detection of mismatching answer behavior has not been addressed so far, and it is probably more challenging, from a computational perspective, than the automatic detection of probe nonresponses, which is based on an empirical approach and clear definitions of probe nonresponses. Although it is feasible that generalizable types of probe nonresponses can be found, it may not be possible to locate instances of probe mismatches, since they may have varied substantive issues. However, previous research already has pointed to the importance of appropriate survey design to prevent instances of mismatches (Behr et al. 2014a). In a similar vein, it also may be very challenging to emulate the interactive situation between a cognitive interviewer and respondents during cognitive interviewing in a web context. It is possible to create a list of “trigger words” that can automatically activate a pre-programmed follow-up probe, but this strategy is only feasible when sufficient knowledge is available about the research topic. For example, it would be possible to ask respondents who misunderstood the term *civil disobedience* an automatic follow-up probe because previous research already has found typical formulations that were used by respondents who misunderstood this key term (Behr et al. 2014b). Given the exploratory character of online probing in most research situations, the potential of imitating interactivity may be limited in this context, which is an evaluation shared by Braun and colleagues (2015): “Comparing typical cognitive interviewing and web probing, it can be stated that items that are difficult to probe and potentially require back-and-forth communication between interviewer and respondent should exclusively be reserved for cognitive interviewing” (p. 195).

The second issue for future research is the *need to reduce the workload* for the coding process of online probing. Given the large sample sizes, the coding process for online probing

is the most work-intensive part of the research process, since a coding schema based on the probe answers has to be developed, and the probes need to be coded and coded a second time for inter-coder reliability. Three potential solutions may help to reduce the workload. First, online probing could use already *established coding schemas*, such as the error coding schema developed by DeMaio and Landreth (2004) or the Question Appraisal System by Willis and Tessler (1999). The application of established coding schemas would enable researchers to skip the development process of the coding schemas. At the same time, it could potentially increase the transferability of results (Willis 2015), since other studies applied the same coding schemas; however, this strategy would limit the analysis to error detection. Second, future research could look into the potential of the *(semi-)automated coding* of probes, which may facilitate the data analyzing process. The success of automated coding, however, depends on the probe types, since they create probe responses of varying complexity. For example, a specific probe may show greater potential for automatic coding than a category-selection probe. Respondents often write lists as an answer to specific probes (e.g., “Italians, Turks, and Chinese” when asked for the type of immigrants they had in mind when answering the question; Behr et al. 2014b). In contrast, the probe responses for a category-selection probe are more complex, since respondents often give reasons and write long responses. As a consequence, specific probes show the greatest potential for (semi-)automated coding. A third alternative, of course, is the *reduction of the sample size* of online probing. Future research should address the question about what constitutes a sufficient sample size for online probing. So far, previous studies have used sample sizes of around 500 respondents for each country, with split conditions reducing the sample size to around 100 respondents for some items. The sub-sample size of 100 respondents could uncover many errors and themes. However, it has not been systematically tested yet whether this sample size could find all the relevant potential problems and associations.

5.2 ARTICLE 2: RESULTS, LIMITATIONS, AND FUTURE RESEARCH

The second goal of online probing—evaluating the cross-national equivalence of survey items—was addressed in the second article (Chapter 3). Online probing shares this research goal with the quantitative approaches of measurement invariance tests, such as MGCFA. Thus, this article answered the following research questions by examining the substantive applications of the constructs of nationalism and constructive patriotism:

1. Do online probing and MGCFA arrive at similar conclusions?
2. Can the insights from online probing help to explain instances of missing comparability?
3. What is the optimal way to combine both methods?

Since previous research indicated that the items measuring constructive patriotism are particularly error prone, this article focused on these three items (“democracy,” “social security,” and “fair and equal”) during online probing.

5.2.1 Results

The results showed that both methods arrive at (partly) similar conclusions. The MGCFA confirmed metric invariance for both constructs, but (partial) scalar measurement invariance tests failed for the five countries in this study. Thus, it is possible to explore structural relationships, but a cross-national comparison of the means of the latent constructs is impossible. The inspection of modification indices and expected parameter changes suggested a cross-loading for the U.S., and an error correlation between the “social security” and “democracy” item for Mexico. The online probing results showed that two of the items (“fair and equal” and “democracy”) were equally understood by the respondents from the five countries. However, the online probe of the item “social security” revealed several problematic issues, such as the varying lexical scope of the term *social security system* across

countries and a silent misinterpretation of this key term by the Mexican respondents. Therefore, both methods agreed that the items are to a certain extent comparable but that some equivalence issues remain. Both methods found that the item “social security” is somewhat problematic. However, the two methods arrived at different conclusions with respect to the item “fair and equal.” MGCFA indicated an increased modification index for the U.S. for this item, whereas online probing revealed a sufficient level of comparability for all countries.

The OP results could also help to explain a finding of the MGCFA. The suggested error correlation between the “social security” and “democracy” items in Mexico was due to the silent misinterpretation of the term *social security system*. Many Mexican respondents understood *security* instead of *social security*. At the same time, the general security situation also was an issue that was revealed in the Mexican probe for the “democracy” item. However, online probing did not find any confirmations that the cross-loading for the U.S., which was suggested by the MGCFA, was due to a lack of equivalence of meaning. These results underline the importance of a theory-driven approach in model improvement in the MGCFA, and prove the potential of online probing for informing and guiding the evaluation of the appropriateness of model modifications.

This article revealed, once again, the complimentary strength of online probing vis-à-vis another evaluation method, which was, in this case, the quantitative MGCFA approach. A combination of both methods can facilitate the equivalence assessment and heighten the insights in an exploratory research situation and the work with more established survey items.

5.2.2 Limitation

A limitation of this study was the impossibility of conducting the MGCFA and online probing with the same data set due to split conditions in the web survey. Although both data sets (the

ISSP data and the web survey) had similar means, standard deviations, and nonresponse rates, an evaluation of both methods with one data set would have been preferable.

5.2.3 Future research

Although the second article showed the potential of a cross-national application of online probing and its combination with MGCFA, several research questions remain to be answered with respect to the cross-national context. Given the increasing globalization of survey research (Heath et al. 2005), more and more countries are included in large-scale comparative survey projects, such as the ISSP or WVS.

An important area for future research concerns questions about an appropriate *sampling strategy* that determines which countries should be included in an online probing study. The selection of countries included in this dissertation was preset, since the data it used was collected within a research project. However, a variety of possible sample strategies exist that should be assessed by future research. The criteria for selection can differ along three lines: convenience, theory driven approaches, and empirically based approaches. The *convenience* approach would select countries on the basis of practical considerations, such as minimizing the effort for translations or having access to experts on specific cultures. The *theory driven* strategy would select countries on the basis of specific theoretical approaches, such as Esping Andersen's welfare regime types (1990) or cultural typologies that were developed by Inglehart and Welzel (2005) or Schwartz (2006) (for a more detailed discussion of country sampling see Boenke et al. 2011). Finally, the *empirically based* strategy would use previous quantitative or qualitative research results to decide on the countries that should be included in a study. The second article of this dissertation already discussed the application of MGCFA to select countries for a cross-national online probing study. However, a variety of other quantitative procedures could facilitate sampling for online probing studies. For example, item response theory (IRT) could indicate problematic countries that have an item

bias (Woehr and Meriac 2010). Online probing could be applied to a sample of countries that contains cases with item bias and cases without item bias to contrast “typical” item interpretations with deviant interpretations that increase the difficulty of answering to a specific item. A similar approach could be applied by using multiple correspondence analysis (MCA) or multidimensional scaling (MDS) (Braun and Johnson 2010). Although the optimal sampling procedure certainly varies with the intended research goal, future research should address this issue more systematically.

The field of measurement invariance tests is constantly evolving. Recently developed were the methods of alignment (Asparouhov and Muthén 2014), exploratory structural equation modelling (Asparouhov and Muthén 2009), and Bayesian structural equation modelling (BSEM) (Muthén and Asparouhov 2012; for a short overview, see Davidov et al. 2014 and for a more extensive discussion see van de Schoot et al. 2015). A combined approach of BSEM and online probing could yield interesting insights. Since MGCFAs tests for scalar measurement invariance failed in many instances, BSEM makes the assumption that these tests are too strict, since they presuppose exact invariance (zero constraints) between countries. BSEM relaxes the exact zero constraints of the tested parameters by substituting them with “approximate” zero constraints, and therefore, allows for some “wiggle room” (van de Schoot et al. 2015). By implementing BSEM, the number of countries showing scalar invariance could be increased in several instances, since many countries failed to achieve exact invariance, but could achieve approximate invariance (e.g., Cieciuch et al. 2014; Zercher et al. 2015). However, so far, established values for appropriate priors or posterior predictive probability values (ppp) do not exist to clearly distinguish when scalar or metric invariance is achieved and when it is not. The question remains as to whether the assumptions of MGCFAs are always too strict or whether, in some instances, some items indeed lack sufficient cross-national equivalence. Similar to the approach of the second article (the comparison of online probing with MGCFAs), online probing could help to distinguish

between instances of “real” measurement invariance and instances where the “wobble room” might have been too large.

5.3 ARTICLE 3: RESULTS, LIMITATIONS, AND FUTURE RESEARCH

The third article of this dissertation (Chapter 4) discussed the potential of online probing as an assessment tool for the cross-national comparability of single-item indicators. The general national pride item served as the substantive application. Additionally, items measuring nationalism and constructive patriotism were used during the exploratory factor analysis to further assess the general national pride item. The third article addressed the following research goals:

1. It assessed the suitability of the general national pride item as a cross-national indicator for the different elements of national identity, and explored the full variety of respondents’ associations when answering this item.
2. It assessed the prevalence of potentially problematic issues.
3. It demonstrated how exploratory factor analysis, online probing, and regression analysis can be combined to detect and explain equivalence issues and to assess the distorting impact of revealed problems.

5.3.1 Results

The exploratory factor analysis (EFA) and the online probing results revealed that respondents associate various concepts with the general national pride item and reject the idea that this item can serve as a proxy for complex concepts such as nationalism or constructive patriotism.

Additionally, the probes uncovered several problematic issues whose distorting impact on answer selection was confirmed by a regression analysis. These results caution against the

use of the GNP item in a cross-national context, but strongly call for a measurement of the different elements of national identity with multiple indicators that are based on a strong theoretical foundation and definition, a perspective that is also shared by Haller (2002):

Thus we must conclude that the use of one single item for a complex dimension like ‘national pride’ is strongly misleading and the analyses based on it can be seriously flawed. It is of utmost importance for international comparative research to ensure the validity and reliability of the indices and scales used. (P. 148)

The combination of EFA, regression analysis, and online probing seems to be an insightful approach to use in an exploratory research situation where already established constructs exist that can be applied during the EFA. The factor structure of EFA can reveal countries that potentially have different understandings of the single-item indicator, and online probing can uncover the reasons for these variations. The regression analysis could provide further insights for evaluating the distorting impact of several problematic issues, such as “missing individual achievement,” “effects of social desirability,” and “missing relevance of national identity as source for pride.” Once again, the different methods are complementary, since they all highlight different aspects of the equivalence issue.

5.3.2 Limitation

Since the third article used two different data sets (2013 ISSP and a web survey), the same limitation as for the second article applied. Due to split conditions, it was impossible to conduct the EFA and online probing with the same data set. Although both data sets had similar means, standard deviations, and nonresponse rates, an evaluation of both methods with one data set would have been preferable.

5.3.3 Future research.

The limitations of previous online probing research also give rise to future research needs. Since until recently no *cross-national probability based web panels* existed, previous online

probing studies had to use non-probability online panels to assess the equivalence of survey items. Fortunately, the situation will change in the next years, as several national probability-based web panels are increasing their international collaborations. Furthermore, the ESS is planning to construct a cross-national probability-based web panel system (work package 7 of the SERISS project; SERISS 2015). An implementation of online probing in the ESS or a similar probability-based web panel could improve the generalizability of online probing results.

Additionally, certain cultures already are known to give answers to closed items that are affected by *response styles*, such as social desirability and acquiescence. For example, more collectivist cultures tend to produce a higher level of social desirability and acquiescence responding than individualistic cultures (Johnson et al. 2011). It is highly probable that these effects also appear in online probing. Future research should control for the impact of response styles on the substantive conclusions of probe answers. At the same time, online probing also can be used to advance the research field of response styles as Johnson (2011) notes:

A general limitation ... is that most of the cross-cultural comparisons of response style measures reviewed have not investigated the role of measurement equivalence. Although this is a general concern, it would appear to be particularly problematic when comparing measures of social desirability across cultures. It will be important for future research to address this oversight. (P. 161)

The third issue for future research is more substantive in nature. As Mohler and Johnson (2010) mentioned, social survey research frequently applies single-item indicators to measures of the social phenomena of interest. Since an equivalence assessment that is based only on quantitative approaches is, in most cases, unfeasible, online probing should be used to *assess further single-item indicators of great relevance*, such as the “top-bottom self-placement” that is an ISSP indicator for subjective class membership.

5.4 CONCLUSION

The three articles of this dissertation support the great potential of online probing for revealing respondents' cognitive processes. They show that online probing is capable of detecting different error sources, uncovering various associations, and assessing the equivalence of constructs and single-item indicators. All three articles also proved the great potential of online probing when it is applied as a complimentary method with qualitative (cognitive interviewing) and quantitative (MGCEFA, EFA, and regression analysis) approaches.

At the same time, these articles found several relevant research questions for future studies, such as interactivity, the reduction of the workload of the coding process of online probing, and the need for strategies for an optimal country sampling. Future research also should address the issue of potential response styles in open-ended questions, and implement online probing in probability-based samples or at least test whether big differences occur in substantive results when a probability-based sample is used instead of a quota-based sample. Future research also should continue to assess the equivalence of further single-item indicators of great relevance. Although these issues were discussed in the context of specific articles, most of the mentioned issues are general future research needs for the application of online probing.

References for Introduction and Conclusion

- Adorno, Theodor W., Else Frenkel-Brunswik, Daniel J. Levinson, and R. Nevitt Sanford. 1950. *The Authoritarian Personality*. Oxford, UK: Harpers.
- Ariely, Gael. 2012. "Globalization, Immigration and National Identity: How the Level of Globalization Affects the Relations between Nationalism, Constructive Patriotism and Attitudes toward Immigrants?" *Group Processes & Intergroup Relations* 15:539–557.
- Ariely, Gael and Eldad Davidov. 2012. "Assessment of Measurement Equivalence with Cross-national and Longitudinal Surveys in Political Science." *European Political Science* 11(3):363–377.
- Asparouhov, Tihomir and Bengt Muthén. 2009. "Exploratory Structural Equation Modeling." *Structural Equation Modeling: A Multidisciplinary Journal* 16(3):397–438.
- Asparouhov, Tihomir and Bengt Muthén. 2014. "Multiple-group Factor Analysis Alignment." *Structural Equation Modeling: A Multidisciplinary Journal* 21(4):495–508.
- Beatty, Paul C. and Gordon B. Willis. 2007. "Research Synthesis: The Practice of Cognitive Interviewing." *Public Opinion Quarterly* 71(2):287–311.
- Bechhofer, Frank and David McCrone. 2010. "Choosing National Identity." *Sociological Research Online* 15(3):3.
- Bechhofer, Frank and David McCrone. 2013. "Imagining the Nation: Symbols of National Culture in England and Scotland." *Ethnicities* 13(5):544–564.
- Behr, Dorothée. 2014. "Translating Answers to Open-ended Survey Questions in Cross-cultural Research: A Case Study on the Interplay between Translation, Coding, and Analysis." *Field Methods* online first:1–16.
doi: <http://dx.doi.org/10.1177/1525822X14553175>.

- Behr, Dorothée, Lars Kaczmirek, Wolfgang Bandilla, and Michael Braun. 2012. "Asking Probing Questions in Web Surveys: Which Factors Have an Impact on the Quality of Responses?" *Social Science Computer Review* 30(4):487–498.
doi: <http://dx.doi.org/10.1177/0894439311435305>.
- Behr, Dorothée, Michael Braun, Lars Kaczmirek, and Wolfgang Bandilla. 2013. "Testing the Validity of Gender Ideology Items by Implementing Probing Questions in Web Surveys." *Field Methods* 25(2):124–141.
doi: <http://dx.doi.org/10.1177/1525822X12462525>.
- Behr, Dorothée, Michael Braun, Lars Kaczmirek, and Wolfgang Bandilla. 2014a. "Item Comparability in Cross-national Surveys: Results from Asking Probing Questions in Cross-national Web Surveys about Attitudes towards Civil Disobedience." *Quality & Quantity* 48(1):127–148. doi: <http://dx.doi.org/10.1007/s11135-012-9754-8>.
- Behr, Dorothée and Michael Braun. 2015. "Satisfaction with the Way Democracy Works: How Respondents across Countries Understand the Question." Pp. 121–138 in *Hopes and Anxieties. Six Waves of the European Social Survey*, edited by P. B. Sztabinski, H. Domanski, and F. Sztabinski. Frankfurt am Main: Lang.
- Behr, Dorothée, Wolfgang Bandilla, Lars Kaczmirek, and Michael Braun. 2014b. "Cognitive Probes in Web Surveys: On the Effect of Different Text Box Size and Probing Exposure on Response Quality." *Social Science Computer Review* 32(4):524–533.
doi: <http://dx.doi.org/10.1177/0894439313485203>.
- Berry, John W. 1969. "On Cross-cultural Comparability." *International Journal of Psychology* 4(2):119–128.
- Bethlehem, Jelke and Silvia Biffignandi. 2012. *Handbook of Web Surveys*. Vol. 567. Hoboken, NJ: Wiley & Sons.

- Blair, Johnny and Frederick G. Conrad. 2011. "Sample Size for Cognitive Interview Pretesting." *Public Opinion Quarterly* 75(4):636–658.
- Blank, Thomas and Peter Schmidt. 2003. "National Identity in a United Germany: Nationalism or Patriotism? An Empirical Test with Representative Data." *Political Psychology* 24:289–311.
- Blasius, Jörg and Victor Thiessen. 2006. "Assessing Data Quality and Construct Comparability in Cross-national Surveys." *European Sociological Review* 22(3): 229–242.
- Boenke, Klaus, Petra Lietz, Margrit Schreier, and Adalbert Wilhelm. 2011. "Sampling: The Selection of Cases for Culturally Comparative Psychological Research." Pp. 101–129 in *Cross-cultural Research Methods in Psychology*, edited by D. Matsumoto and F. van de Vijver. New York, NY: Cambridge University Press.
- Bollen, Kenneth. 1989. *Structural Equations with Latent Variables*. New York, NY: John Wiley & Sons.
- Bollen, Kenneth and Juan Diez Medrano. 1998. "Who Are the Spaniards? Nationalism and Identification in Spain." *Social Forces* 77(2):587–621.
- Bonikowski, Bart. 2009. "Beyond National Identity: Collective Schemata of the Nation in Thirty Countries." *APSA 2009 Toronto Meeting Paper*.
- Braun, Michael. 2003. "Communication and Social Cognition." Pp. 57–68 in *Cross-cultural Survey Methods*, edited by J. Harkness, F. van de Vijver, and P. Mohler. Hoboken, NJ: Wiley-Interscience.
- Braun, Michael, Dorothée Behr, and Lars Kaczmirek. 2013. "Assessing Cross-national Equivalence of Measures of Xenophobia: Evidence from Probing in Web Surveys." *International Journal of Public Opinion Research* 25(3):383–395.
doi: <http://dx.doi.org/10.1093/ijpor/eds034>.

- Braun, Michael, Dorothée Behr, Lars Kaczmarek, and Wolfgang Bandilla. 2015. "Evaluating Cross-national Item Equivalence with Probing Questions in Web Surveys." Pp. 184–200 in *Improving Survey Methods: Lessons from Recent Research*, edited by U. Engel, B. Jann, P. Lynn, A. Scherpenzeel, and P. Sturgis. New York, NY: Routledge.
- Braun, Michael and Timothy Johnson. 2010. "An Illustrative Review of Techniques for Detecting Inequivalences." Pp. 373–393 in *Survey Methods in Multinational, Multiregional, and Multicultural Contexts*, edited by J. A. Harkness, M. Braun, B. Edwards, T. P. Johnson, L. Lyberg, P. Ph. Mohler, B. Pennell, and T. W. Smith. Hoboken, NJ: Wiley-Blackwell.
- Braun, Michael and Jaqueline Scott. 1998. "Multidimensional Scaling and Equivalence: Is Having a Job the Same as Working?" Pp. 129–144 in *Zuma-Nachrichten Spezial*, Vol. 3, *Cross-cultural Survey Equivalence*, edited by J. Harkness. Mannheim: Zuma.
- Braun, Michael and Rolf 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 J. H. Hoffmeyer-Zlotnik and C. Wolf. New York, NY: Kluwer Academic/Plenum Publishers.
- Caramelli, Marco and Fons van de Vijver. 2013. "Towards a Comprehensive Procedure for Developing Measurement Scales for Cross-cultural Management Research." *Management International/International Management/Gestión Internacional* 17(2):150–163.
- Cieciuch, Jan, Eldad Davidov, Peter Schmidt, René Algesheimer, and Shalom H. Schwartz. 2014. "Comparing Results of an Exact vs. an Approximate (Bayesian) Measurement Invariance Test: A Cross-country Illustration with a Scale to Measure 19 Human Values." *Frontiers in Psychology* 5(982):1–10.
- Cohen, Robin. 1994. *Frontiers of Identity: The British and the Others*. London, UK: Longman.

- Collins, Debbie. 2015. "Cognitive Interviewing: Origin, Purpose, and Limitations." Pp. 1–27 in *Cognitive Interviewing Practice*, edited by D. Collins. London, UK: Sage
- Davidov, Eldad. 2009. "Measurement Equivalence of Nationalism and Constructive Patriotism in the ISSP: 34 Countries in a Comparative Perspective." *Political Analysis* 17(1):64–82.
- Davidov, Eldad. 2011. "Nationalism and Constructive Patriotism: A Longitudinal Test of Comparability in 22 Countries with the ISSP." *International Journal of Public Opinion Research* 23(1):88–103.
- Davidov, Eldad, Hermann Dülmer, Elmar Schlüter, Peter Schmidt, and Bart Meuleman. 2012. "Using a Multilevel Structural Equation Modeling Approach to Explain Cross-cultural Measurement Noninvariance." *Journal of Cross-cultural Psychology* 43(4):558–575.
- Davidov, Eldad, Bart Meuleman, Jan Cieciuch, Peter Schmidt, and Jaak Billiet. 2014. "Measurement Equivalence in Cross-national Research." *Annual Review of Sociology* 40:55–75.
- DeMaio, Theresa J. and Ashley Landreth. 2004. "Do Different Cognitive Interview Techniques Produce Different Results?" Pp. 89–108 in *Methods for Testing and Evaluating Survey Questionnaires*, edited by S. Presser, J. M. Rothgeb, M. P. Couper, J. T. Lessler, E. Martin, J. Martin, and E. Singer. Hoboken, NJ: John Wiley & Sons.
- Denzin, Norman K. 1970. "Triangulation: A case for Methodological Evaluation and Combination." Pp. 339–357 in *Sociological Methods: A Source Book*, edited by N.K. Denzing. Chicago: Aldine.
- Elkins, Zachary and John Sides. 2006. *In Search of the Unified Nation-state: National Attachment among Distinctive Citizens*. Irvine, CA: Center for the Study of Democracy. Retrieved October 13, 2015 (<https://escholarship.org/uc/item/20f203bx>).
- Encarnación, Omar G. 2009. "Spain Remade, Again." *Current History* 108(716):117–123.

- Esping-Andersen, Gøsta. 1990. *The Three Worlds of Welfare Capitalism*. Cambridge, UK: Polity Press.
- Fitzgerald, Rory, Sally Widdop, S., Michelle Gray, and Debbie Collins. 2009. *Testing for Equivalence Using Cross-national Cognitive Interviewing*. London, UK: Centre for Comparative Social Surveys. Retrieved October 10, 2015 (https://www.city.ac.uk/__data/assets/pdf_file/0014/125132/CCSS-Working-Paper-No-01.pdf).
- Fleiß, Jürgen, Franz Höllinger, and Helmut Kuzmics. 2009. "Nationalstolz zwischen Patriotismus und Nationalismus?" *Berliner Journal für Soziologie* 19(3):409–434.
- Fontaine, Johnny R. J. 2005. "Equivalence." Pp. 803–813 in *Encyclopedia of Social Measurement*, edited by K. Kempf-Leonard. Amsterdam, Netherlands: Elsevier Academic Press.
- Freedom House. 2015. *Freedom in the World 2014: The Annual Survey of Political Rights and Civil Liberties*. New York, NY: Rowman & Littlefield.
- Fujishiro, Kaori, Fang Gong, Sherry Baron, C. Jeffery Jacobson, Sheli DeLaney, Michael Flynn, and Donald E. Eggerth. 2010. "Translating Questionnaire Items for a Multilingual Worker Population: The Iterative Process of Translation and Cognitive Interviews with English-, Spanish-, and Chinese-Speaking Workers." *American Journal of Industrial Medicine* 53(2):194–203.
- GESIS. 2015. *GESIS-Variable Reports No. 2015/35 ISSP 2013 National Identity III Variable Report*. Köln: GESIS Data Archive for the Social Sciences. Retrieved November 12, 2015 (https://dbk.gesis.org/dbksearch/download.asp?file=ZA5950_cdb.pdf).
- Gonzalez, Francisco. 2012. *Freedom House—Countries at the Crossroads 2012: Mexico*. Washington, DC: Freedom House. Retrieved May 20, 2015 (<https://freedomhouse.org/report/countries-crossroads/2012/mexico>).

- Gray, Michelle and Margaret Blake. 2015. "Cross-national, Cross-cultural and Multilingual Cognitive Interviewing." Pp. 220–242 in *Cognitive Interviewing Practice*, edited by Debbie Collins. London, UK: Sage Publications.
- Haller, Max. 2002. "Theory and Method in the Comparative Study of Values: Critique and Alternative to Inglehart." *European Sociological Review* 18(2):139–158.
- Haller, Max. 2009. "Introduction." Pp. 172–174 in *The International Social Survey Programme 1984–2009: Charting the Globe*, edited by M. Haller, R. Jowell, and T. W. Smith. London and New York: Routledge.
- Harkness, Janet. 2008. "Comparative Survey Research: Goals and Challenges." Pp. 56–77 in *International Handbook of Survey Methodology*, edited by E. De Leeuw, D. Edith, and D. Dillman. New York and London: Taylor & Francis.
- Harkness, Janet, Brad Edwards, Sue Ellen Hansen, Debra R. Miller, and Ana Villar. 2010. "Designing Questionnaires for Multipopulation Research." Pp. 31–57 in *Survey Methods in Multinational, Multiregional, and Multicultural Contexts*, edited by J. A. Harkness, M. Braun, B. Edwards, T. P. Johnson, L. Lyberg, P. Ph. Mohler, B. Pennell, and T. W. Smith. Hoboken, NJ: Wiley-Blackwell.
- He, Jia and Fons van de Vijver. 2012. "Bias and Equivalence in Cross-cultural Research." *Online Readings in Psychology and Culture* 2(2):1–18
- Heath, Anthony, Stephen Fisher, and Shawna Smith. 2005. "The Globalization of Public Opinion Research." *Annual Review of Political Science* 8: 297–333.
- Heath, Anthony, Jean Martin, and Thees Spreckelsen. 2009. "Cross-national Comparability of Survey Attitude Measures." *International Journal of Public Opinion Research* 21(3):293–315.

- Heeringa, Steven G. and Colm O'muircheartaigh. 2010. "Sample Design for Cross-cultural and Cross-national Survey Programs." Pp. 251–267 in *Survey Methods in Multinational, Multiregional, and Multicultural Contexts*, edited by J. A. Harkness, M. Braun, B. Edwards, T. P. Johnson, L. Lyberg, P. Ph. Mohler, B. Pennell, and T. W. Smith. Hoboken, NJ: Wiley-Blackwell.
- Hjerm, Mikael. 1998. "National Identities, National Pride and Xenophobia: A Comparison of Four Western Countries." *Acta Sociologica* 41(4):335–347.
- Hoffmeyer-Zlotnik, Jürgen H. and Janet A. Harkness. 2005. "Methodological Aspects in Cross-national Research: Foreword." Pp. 5–10 in *ZUMA Nachrichten*, Vol. 11, *Methodological Aspects in Cross-national Research*, edited by J. H. Hoffmeyer-Zlotnik and J. A. Harkness. Mannheim: Zuma.
- Holland, Jennifer L. and Leah Melani Christian. 2009. "The Influence of Topic Interest and Interactive Probing and Responses to Open-ended Questions in Web Surveys." *Social Science Computer Review* 27(2):196–212.
- Horn, John L. and Jack J. McArdle. 1992. "A Practical and Theoretical Guide to Measurement Invariance in Aging Research." *Experimental Aging Research* 18(3):117–144.
- Huddy, Leonie and Nadia Khatib. 2007. "American Patriotism, National Identity, and Political Involvement." *American Journal of Political Science* 51(1):63–77.
- Hui, C. Harry and Harry C. Triandis. 1985. "Measurement in Cross-cultural Psychology: A Review and Comparison of Strategies." *Journal of Cross-cultural Psychology* 16(2):131–52.
- Hutcheson, John, David Domk, Andre Billeaudeau, and Philip Garland. 2004. "US National Identity, Political Elites, and a Patriotic Press Following September 11." *Political Communication*, 21(1):27–50.

- Inglehart, Ronald and Christian Welzel. 2005. *Modernization, Cultural Change, and Democracy: The Human Development Sequence*. Cambridge, UK: Cambridge University Press.
- ISSP. 2015a. *International Social Survey Programme—Members and Addresses*. Retrieved November 27, 2015 (<http://www.issp.org/page.php?pageId=2>).
- ISSP. 2015b. *International Social Survey Programme—General Information: History of the ISSP*. Retrieved November 27, 2015 (<http://www.issp.org/page.php?pageId=216>).
- Johnson, Timothy P. 1998. “Approaches to Equivalence in Cross-cultural and Cross-national Survey Research.” Pp. 1–40 in *Zuma-Nachrichten Spezial*, Vol. 3, *Cross-cultural Survey Equivalence*, edited by J. A. Harkness. Mannheim: Zuma.
- Johnson, Timothy P., Sharon Shavitt, and Allyson Holbrook. 2011. “Survey Response Styles across Cultures.” Pp. 130–175 in *Cross-cultural Research Methods in Psychology*, edited by D. Matsumoto and F. van de Vijver. New York, NY: Cambridge University Press.
- Jöreskog, Karl G. 1971. “Simultaneous Factor Analysis in Several Populations.” *Psychometrika* 36(4):409–426.
- Kaczmarek, Lars, Katharina Meitinger, and Dorothee Behr. 2015. “Item Nonresponse in Open-ended Questions: Identification and Reduction in Web Surveys.” Presentation at *ESRA 2015: 6th Conference of the European Survey Research Association*, Reykjavik. Retrieved November 20, 2015 (<http://www.europeansurveyresearch.org/conference/programme2015?sess=33>).
- Kankaraš, Miloš, Jeroen K. Vermunt, and Guy Moors. 2011. “Measurement Equivalence of Ordinal Items: A Comparison of Factor Analytic, Item Response Theory, and Latent Class Approaches.” *Sociological Methods & Research* 40(2):279–310.
- Kosterman, Rick and Seymour Feshbach. 1989. “Toward a Measure of Patriotic and Nationalistic Attitudes.” *Political Psychology* 10(2):257–274.

- Kelley, Jonathan and M. D. R. Evans. 2002. "National Pride in the Developed World." *International Journal of Public Opinion Research* 14(3):1–24.
- Krosnick, Jon A. 1991. "Response Strategies for Coping with the Cognitive Demands of Attitude Measures in Surveys." *Applied Cognitive Psychology* 5(3): 213–236.
- Latcheva, Rossalina. 2011. "Cognitive Interviewing and Factor-analytic Techniques: A Mixed Method Approach to Validity of Survey Items Measuring National Identity." *Quality & Quantity* 45(6):1175–1199.
- Liu, Lu, Mandy Sha, and Hyunjoo Park. 2013. "Exploring the Efficiency and Utility of Methods to Recruit Non-English Speaking Qualitative Research Participants." *Survey Practice* 6(3):1–8.
- Lynn, Peter, Lilli Japac, and Lars Lyberg. 2006. "What's So Special about Cross-national Surveys?" Pp. 7–21 in *Zuma-Nachrichten Spezial*, Vol. 12, *Conducting Cross-national and Cross-cultural Surveys*, edited by J. Harkness. Mannheim: Zuma.
- MacInnes, John. 2006. "Category and Comparison across What Kind of Frontier?" Pp. 101–114 in *Zuma-Nachrichten Spezial*, Vol. 12, *Conducting Cross-national and Cross-cultural Surveys*, edited by J. Harkness. Mannheim: Zuma.
- McCrone, David. 1998. *The Sociology of Nationalism: Tomorrow's Ancestors*. London and New York: Routledge.
- McCrone, David. 2002. "Who Do You Say You Are? Making Sense of National Identities in Modern Britain." *Ethnicities* 2(3):301–320.
- Medina, Tait R., Shawna Smith, and J. Scott Long. 2009. "Measurement Models Matter: Implicit Assumptions and Cross-national Research." *International Journal of Public Opinion Research* 21(3):333–361.
- Medrano, Juan Díez and Paula Gutiérrez. 2001. "Nested Identities: National and European Identity in Spain." *Ethnic and Racial Studies* 24(5):753–778.

- Meredith, William. 1964. "Rotation to Achieve Factorial Invariance." *Psychometrika* 29(2):187–206.
- Meredith, William. 1993. "Measurement Invariance, Factor Analysis and Factorial Invariance." *Psychometrika* 58(4):525–543.
- Merino, Stephen M. 2010. "Religious Diversity in a 'Christian Nation': The Effects of Theological Exclusivity and Interreligious Contact on the Acceptance of Religious Diversity." *Journal for the Scientific Study of Religion* 49(2):231–246.
- Miller, David and Sundas Ali. 2014. "Testing the National Identity Argument." *European Political Science Review* 6(2):237–259.
- Miller, Kirsten, Daniel Mont, Aaron Maitland, Barbara Altman, and Jennifer Madans. 2011. "Results of a Cross-national Structured Cognitive Interviewing Protocol to Test Measures of Disability." *Quality & Quantity* 45(4):801–815.
- Miller-Idriss, Cynthia. 2009. *Blood and Culture: Youth, Right-wing Extremism, and National Belonging in Contemporary Germany*. Durham: Duke University Press.
- Moghaddam, Fathali M., Benjamin R. Walker, and Rom Harre. 2003. "Cultural Distance, Levels of Abstraction, and the Advantages of Mixed Methods." Pp.111–134 in *Handbook of Mixed Methods in Social and Behavioral Research*, edited by A. Tashakkori and C. Teddlie. Thousand Oaks, CA: Sage.
- Mohler, Peter and Timothy Johnson. 2010. "Equivalence, Comparability, and Methodological Progress." Pp. 17–29 in *Survey Methods in Multinational, Multiregional, and Multicultural Contexts*, edited by J. A. Harkness, M. Braun, B. Edwards, T. P. Johnson, L. Lyberg, P. Ph. Mohler, B. Pennell, and T. W. Smith. Hoboken, NJ: Wiley-Blackwell.
- Morris, Stephen D. 1999. "Reforming the Nation: Mexican Nationalism in Context." *Journal of Latin American Studies* 31(02):363–397.

- Muñoz, Jordi. 2009. "From National-Catholicism to Democratic Patriotism? Democratization and Reconstruction of National Pride: The Case of Spain (1981–2000)." *Ethnic and Racial Studies* 32(4):616–639.
- Muthén, Bengt and Tihomir Asparouhov. 2012. "Bayesian Structural Equation Modeling: A More Flexible Representation of Substantive Theory." *Psychological Methods* 17(3): 313–335.
- Prüfer, Peter and Margrit Rexroth. 2005. "Kognitive Interviews." *ZUMA How-to-Reihe 15*. Retrieved October 20, 2015 (http://www.gesis.org/fileadmin/upload/forschung/publikationen/gesis_reihen/howto/How_to15PP_MR.pdf).
- Reeskens, Tim and Matthew Wright. 2013. "Nationalism and the Cohesive Society a Multilevel Analysis of the Interplay among Diversity, National Identity, and Social Capital across 27 European Societies." *Comparative Political Studies* 46(2):153–181.
- Schatz, Robert T., Ervin Staub, and Howard Lavine. 1999. "On the Varieties of National Attachment: Blind versus Constructive Patriotism." *Political Psychology* 20(1):151–174.
- Schildkraut, Deborah J. 2014. "Boundaries of American Identity: Evolving Understandings of 'Us.'" *Annual Review of Political Science* 17:441–460.
- Schwartz, Shalom H. 2006. "A Theory of Cultural Value Orientations: Explication and Applications." *Comparative Sociology* 5(2):137–182.
- SERISS. 2015. "WP7: A Survey Future Online" London, UK: European Social Survey. Retrieved November 20, 2015 (<http://seriss.eu/about-seriss/work-packages/wp7-a-survey-future-online/>).
- Sinnott, Richard. 2006. "An Evaluation of the Measurement of National, Subnational and Supranational Identity in Cross-national Surveys." *International Journal of Public Opinion Research* 18(2):211–223.

- Skjåk, Knut Kalgraff. 2010 “The International Social Survey Programme: Annual Cross-national Social Survey Since 1985.” Pp. 497–506 in *Survey Methods in Multinational, Multiregional, and Multicultural Contexts*, edited by J. A. Harkness, M. Braun, B. Edwards, T. P. Johnson, L. Lyberg, P. Ph. Mohler, B. Pennell, and T. W. Smith. Hoboken, NJ: Wiley-Blackwell.
- Smith, Anthony D. 1991. *National Identity*. London, UK: Penguin Books.
- Smith, Shawna, Stephen Fisher, and Anthony Heath. 2011. “Opportunities and Challenges in the Expansion of Cross-national Survey Research.” *International Journal of Social Research Methodology* 14(6):485–502.
- Smith, Tom W. 2010. “The Globalization of Survey Research.” Pp. 475–484 in *Survey Methods in Multinational, Multiregional, and Multicultural Contexts*, edited by J. A. Harkness, M. Braun, B. Edwards, T. P. Johnson, L. Lyberg, P. Ph. Mohler, B. Pennell, and T. W. Smith. Hoboken, NJ: Wiley-Blackwell.
- Smith, Tom W. and Lars Jarkko. 2001. *National Pride in Cross-national Perspective*. Chicago, IL: National Opinion Research Center. Retrieved October 20, 2015 (www.issp.org/documents/natpride.doc).
- Smith, Tom W. and Seokho Kim. 2006. “National Pride in Comparative Perspective: 1995/96 and 2003/04.” *International Journal of Public Opinion Research* 18(1):127–136.
- Smith, Tom W. 2009. “The ISSP: History, Organization and Members, Working Principles and Outcomes. An Historical-Sociological Account.” Pp. 2–28 in *The International Social Survey Programme 1984–2009: Charting the Globe*, edited by M. Haller, R. Jowell, and T. W. Smith. London and New York: Routledge.
- Smith, Tom W. and Benjamin Schapiro. 2015. *A Compilation of Documents Used to Develop the National Identity III Questionnaire for the International Social Survey Programme in 2014*. Chicago, IL: NORC. Retrieved October 20, 2015 (<http://www.esis.org/issp/modules/issp-modules-by-topic/national-identity/2013/>).

- Smyth, Jolene D., Don A. Dillman, Leah M. Christian, and Mallory McBride. 2009. "Open-ended Questions in Web Surveys: Can Increasing the Size of Answer Boxes and Providing Extra Verbal Instructions Improve Response Quality?" *Public Opinion Quarterly* 73:325–337.
- Solt, Frederick. 2011. "Diversionary Nationalism: Economic Inequality and the Formation of National Pride." *The Journal of Politics* 73(3):821–830.
- Steenkamp, Jan-Benedict E. and Hans Baumgartner. 1998. "Assessing Measurement Invariance in Cross-national Consumer Research." *Journal of Consumer Research* 25(1):78–107.
- Tajfel, Henri and John C. Turner. 1986. "The Social Identity Theory of Intergroup Behavior." Pp. 7–24 in *Psychology of Intergroup Relations*, edited by W. G. Austin and S. Worchel. Chicago, IL: Nelson-Hall.
- Tilley, James and Anthony Heath. 2007. "The Decline of British National Pride." *The British Journal of Sociology* 58(4):661–678.
- Theiss-Morse, Elizabeth and Sergio Wals. 2014. "Con La Vara que Midas... National Identity and Attitudes toward Immigration and Emigration in Mexico." Paper presented at the Annual Meeting of the American Political Science Association, Washington, DC, August 28–31, 2014.
- Torcal, Mariano. 2014. "The Decline of Political Trust in Spain and Portugal Economic Performance or Political Responsiveness." *American Behavioral Scientist* 58(12):1542–1567.
- Tourangeau, Roger and Kenneth A. Rasinski. 1988. "Cognitive Processes Underlying Context Effects in Attitude Measurement." *Psychological Bulletin* 103(3):299.
- Tourangeau, Roger, Lance J. Rips, and Kenneth Rasinski. 2000. *The psychology of Survey Response*. Cambridge, UK: Cambridge University Press.

- Van De Schoot, Rens, Peter Schmidt, Alain De Beuckelaer, Kimberley Lek, and Marielle Zondervan-Zwijnenburg. 2015 "Editorial: Measurement Invariance." *Frontiers in Psychology* 6:1–4.
- Van de Vijver, Fons J. 2011. "Capturing Bias in Structural Equation Modeling." Pp. 3–34 in *Cross-cultural Analysis: Methods, Theories, and Empirical Applications in the Social Sciences*, edited by S. Salzborn, E. Davidov, and J. Reinecke. Wiesbaden: VS Verlag für Sozialwissenschaften.
- Van de Vijver, Fons and Athanasios Chasiotis. 2010. "Making Methods Meet: Mixed Designs in Cross-Cultural Research." Pp. 455–473 in *Survey Methods in Multinational, Multiregional, and Multicultural Contexts*, edited by J. A. Harkness, M. Braun, B. Edwards, T. P. Johnson, L. Lyberg, P. Ph. Mohler, B. Pennell, and T. W. Smith. Hoboken, NJ: Wiley-Blackwell.
- Van de Vijver, Fons and Kwok Leung. 1997. *Methods and Data Analysis for Cross-cultural Research*. Thousand Oaks, CA: Sage.
- Van de Vijver and Kwok Leung. 2011. "Equivalence and Bias: A Review of Concepts, Models, Data Analytic Procedures" Pp. 17–45 in *Cross-cultural Research Methods in Psychology*, edited by D. Matsumoto and F. van de Vijver. New York, NY: Cambridge University Press.
- Van de Vijver, Fons and Ype H. Poortinga. 1997. "Towards an Integrated Analysis of Bias in Cross-cultural Assessment." *European Journal of Psychological Assessment* 13(1):29.
- Van de Vijver, Fons and Norbert K. Tanzer. 2004. "Bias and Equivalence in Cross-cultural Assessment: An Overview." *Revue Européenne de Psychologie Appliquée/European Review of Applied Psychology* 54(2):119–135.
- Vandenberg, Robert J. and Charles E. Lance. 2000. "A Review and Synthesis of the Measurement Invariance Literature: Suggestions, Practices, and Recommendations for Organizational Research." *Organizational Research Methods* 3(1):4–70.

- Willis, Gordon B. 2005. *Cognitive Interviewing: A Tool for Improving Questionnaire Design*. Thousand Oaks, CA: Sage Publications.
- Willis, Gordon B. 2015. "The Practice of Cross-cultural Cognitive Interviewing." *Public Opinion Quarterly* 79(1):359–395.
- Willis, Gordon B. and Judith T. Lessler. 1999. *Question Appraisal System QAS-99*. Rockville, MD: Research Triangle Institute.
- Woehr, David J. and John P. Meriac. 2010. "Using Polytomous Item Response Theory to Examine Differential Item and Test Functioning: The Case of Work Ethic." Pp. 419–433 in *Survey Methods in Multinational, Multiregional, and Multicultural Contexts*, edited by J. A. Harkness, M. Braun, B. Edwards, T. P. Johnson, L. Lyberg, P. Ph. Mohler, B. Pennell, and T. W. Smith. Hoboken, NJ: Wiley-Blackwell.
- Zercher, Florian, Peter Schmidt, Jan Ciecuch, and Eldad Davidov. 2015. "The Comparability of the Universalism Value over Time and across Countries in the European Social Survey: Exact versus Approximate Measurement Invariance." *Frontiers in Psychology* 6(733):1–11.

Eidesstattliche Erklärung

Eidesstattliche Versicherung gemäß § 9 Absatz 1 Buchstabe e) der Promotionsordnung der Universität Mannheim zur Erlangung des Doktorgrades der Sozialwissenschaften:

1. Bei der eingereichten Dissertation mit dem Titel „Searching for Equivalence. An Exploration of the Potential of Online Probing with Examples from National Identity“ handelt es sich um mein eigenständig erstelltes eigenes Werk.
2. Ich habe nur die angegebenen Quellen und Hilfsmittel benutzt und mich keiner unzulässigen Hilfe Dritter bedient. Insbesondere habe ich wörtliche Zitate aus anderen Werken als solche kenntlich gemacht.
3. Die Arbeit oder Teile davon habe ich bisher nicht an einer Hochschule des In- oder Auslands als Bestandteil einer Prüfungs- oder Qualifikationsleistung vorgelegt.
4. Die Richtigkeit der vorstehenden Erklärung bestätige ich.
5. Die Bedeutung der eidesstattlichen Versicherung und die strafrechtlichen Folgen einer unrichtigen und unvollständigen eidesstattlichen Versicherung sind mir bekannt.

Ich versichere an Eides statt, dass ich nach bestem Wissen die reine Wahrheit erklärt und nichts verschwiegen habe.

Ort und Datum

Katharina Meitinger