mda
methods, data, analyses
Journal for Quantitative Methods and Survey Methodology
Volume 14, 2020 | 1

methods, data, analyses is published by GESIS – Leibniz Institute for the Social Sciences.

Managing editor: Sabine Häder
GESIS – Leibniz Institute for the Social Sciences
PO Box 12 21 55
68072 Mannheim
Germany
Tel.: + 49.621.1246526
E-mail: mda@gesis.org
Internet: www.mda.gesis.org

Methods, data, analyses (mda) publishes research on all questions important to quantitative methods, with a special emphasis on survey methodology. In spite of this focus we welcome contributions on other methodological aspects. We especially invite authors to submit articles extending the profession‘s knowledge on the science of surveys, be it on data collection, measurement, or data analysis and statistics. We also welcome applied papers that deal with the use of quantitative methods in practice, with teaching quantitative methods, or that present the use of a particular state-of-the-art method using an example for illustration.
All papers submitted to mda will first be screened by the editors for general suitability and then double-blindly reviewed by at least two reviewers. Mda appears in two regular issues per year (January and July).

Layout: Bettina Zacharias (GESIS)

ISSN 1864-6956 (Print)
ISSN 2190-4936 (Online)

Content

RESEARCH REPORTS

methods, data, analyses | Vol. 14(1), 2020, pp. 3-34 DOI: 10.12758/mda.2019.05

Effects of Respondent and Survey Characteristics on the Response Quality of an Open-Ended Attitude Question in Web Surveys

Katharina Schmidt, Tobias Gummer & Joss Roßmann
GESIS – Leibniz Institute for the Social Sciences

Abstract

Open-ended questions have a great potential for analyses, but answering them often imposes a great burden on respondents. Relying on satisficing theory as an overarching theoretical framework, we derived several hypotheses about how respondent and survey level characteristics, and their interactions, might affect the quality of the responses to an open-ended attitude question in self-administered surveys. By applying multilevel analyses to data from 29 web surveys, we examined the effects of respondent and survey level characteristics on three indicators of response quality: response length, response latency, and the interpretability of the answers. With respect to all three indicators, we found that more educated and more motivated respondents provided answers of significantly better quality compared to other respondents. However, the present study provides evidence that analyzing response quality exclusively with process-generated measures of quality may produce a misleading picture. Therefore, the addition of content-related indicators, such as the interpretability of responses, provides a more informative result. We found that the further the open-ended question was located towards the end of the questionnaire, the fewer interpretable answers were given. Our results also indicated that if the survey was carried out in close proximity to a federal election, responses were more likely to be interpretable. Overall, our study suggests that the characteristics at the respondent and survey levels influence the response quality of open-ended attitude questions and that these characteristics interact to a small degree.

Keywords: Open-ended questions, response quality, web surveys, multilevel modeling, satisficing

Researchers make use of open-ended questions in surveys because they allow respondents to report facts, behaviors, or attitudes without being restricted to a fixed set of answer choices. Open-ended questions can produce a much more diverse set of answers compared to closed-ended questions, which influence respondents' answers by providing cues to what kind of information is being sought via their response format (Dillman, Smyth, & Christian, 2014; Fuchs, 2009; Reja, Manfreda, Hlebec, & Vehovar, 2003; Schuman & Presser, 1996; Tourangeau, Rips, & Rasinski, 2000). Although it is well established that open-ended questions are advantageous because researchers can collect rich and detailed information from respondents on a topic of interest, these questions often suffer from comparably lower response quality as, for instance, is indicated by higher levels of item nonresponse (Reja et al., 2003; Schuman & Presser, 1996).

However, the implications for response quality also depend on the type of open-ended questions. Factual or behavioral open-ended questions – for instance, questions on behavioral frequencies or personal characteristics (cf. Fuchs, 2009; Holbrook et al., 2014) – usually limit the universe of adequate responses because the requested form of answers is rather obvious. Yet, particularly open-ended questions that ask about frequencies often suffer from the problem that respondents provide rounded answers (cf. Holbrook et al., 2014; Tourangeau et al., 2000; Turner, Sturgis, & Martin, 2015). Answering open-ended attitude questions is usually more demanding for respondents because they ask for a detailed response that might include several themes and elaboration on these themes (Holland & Christian, 2009; Smyth, Dillman, Christian, & McBride, 2009). Thus, responding to open-ended attitude questions often requires substantial cognitive effort from respondents, which is more burdensome and can lead to respondent fatigue (Dillman et al., 2014; Gummer & Roßmann, 2015; Holland & Christian, 2009). Consequently, respondents may use *satisficing* response strategies to reduce the burden of answering cognitively demanding open-ended attitude questions, which results in answers of lower quality (Holland & Christian, 2009; Krosnick, 1991, 1999).

The susceptibility of open-ended attitude questions to satisficing response behavior is particularly relevant for self-administered surveys, which lack a human interviewer who can motivate respondents and guide them through the response process (Holland & Christian, 2009; Rada & Dominguez-Alvarez, 2013; Reja et al., 2003). A substantial body of methodological research has examined the effects of questionnaire and question design on response quality for web surveys (e.g., Couper, Tourangeau, Conrad, & Zhang, 2013; Smyth et al., 2009; Tourangeau, Couper, & Conrad, 2004). With regard to these considerations, the present study

Direct correspondence to
Katharina Schmidt, GESIS – Leibniz Institute for the Social Sciences
katharina.schmidt@gesis.org

aims to answer the following research question: What characteristics affect the response quality of open-ended attitude questions in web surveys? Identifying relevant characteristics at the respondent and survey levels should effectively support researchers in designing web surveys that generate high-quality responses to open-ended questions.

Previous studies have compared the response quality of open-ended and closed questions (e.g., Reja et al., 2003), or have examined the mode differences in the response quality of web and paper questionnaires with respect to open-ended questions (e.g., Kwak & Radler, 2002; Rada & Dominguez-Alvarez, 2013). In addition, existing research has mostly examined the effects of a limited number of characteristics on response quality, such as the interest of the respondent in the topic (e.g., Galesic & Bosnjak, 2009; Holland & Christian, 2009; Olson & Peytchev, 2007), mobile device usage (Revilla & Ochoa, 2015a; Toepoel & Lugtig, 2014), and gender, age, or education (Couper & Kreuter, 2013; Denscombe, 2007; Yan & Tourangeau, 2008). The present study complements these studies in at least two ways. First, by applying multilevel modeling to data from 29 web surveys, we examined the characteristics of response quality at the respondent and survey levels, and the interaction of the variables at both levels. Second, with the notable exception of Holland and Christian (2009) and Smyth et al. (2009), prior studies have mostly used response length (e.g., Galesic, 2006; Galesic & Bosnjak, 2009; Grauenhorst, Blohm, & Koch, 2016; Kwak & Radler, 2002; Mavletova, 2013; Rada & Dominguez-Alvarez, 2013) or response time (e.g., Callegaro, Yang, Bhola, & Dillman, 2004; Galesic & Bosnjak, 2009) as indicators of response quality. The present study extends this research by using the interpretability of the responses to open-ended questions as an additional indicator of quality. As we argue later in the study, the interpretability of responses is potentially an even more appropriate and informative indicator of response quality than response length or latency.

The remainder of this study is organized as follows. The next section introduces *satisficing* as the theoretical framework for our study. Therefore, we present our expectations on how respondents cope with the cognitive demands of open-ended attitude questions, and review the indicators of response quality that previous research has used. Then, by using satisficing theory, we derive a set of hypotheses on the effects of several survey and respondent characteristics on the response quality of open-ended attitude questions. The following sections describe the data, the operationalization of the independent and dependent variables, and the methods used in the empirical analysis. The last sections present and discuss the results and close with recommendations for further research.

Theoretical Background

In the present study, we use satisficing theory (Krosnick, 1991, 1999) to measure and explain the response quality of open-ended attitude questions. Satisficing theory provides theoretical mechanisms that link the characteristics of questions and respondents with the use of response strategies that negatively affect response quality.

Satisficing theory assumes that answering survey questions usually requires respondents to pass through four stages of cognitive processing (Tourangeau & Rasinski, 1988; Tourangeau et al., 2000) – comprehension, information retrieval, judgment, and response selection. The response strategy that involves the complete and effortful execution of these cognitive processes is termed *optimizing*. However, if the difficulty of a question is high and a respondent is low in ability and/or motivation, the respondent might decide to use a *satisficing* response strategy (Krosnick, 1991, 1999). While weak forms of satisficing imply less effortful cognitive processing, strong satisficing involves skipping altogether the cognitive processes of question comprehension, information retrieval, and judgment. Hence, satisficing enables respondents to reduce the burden of responding to cognitively demanding survey questions (Krosnick, 1991, 1999). Consequentially, it follows from the propositions of the satisficing framework that the quality of responses should be poorer when respondents adopt weak or strong satisficing than when they optimize.

Coping with the Cognitive Demands of Open-Ended Attitude Questions

Satisficing theory states that under the condition of weak satisficing, respondents superficially or incompletely execute the processes of question comprehension, information retrieval, integration of the information into a summarizing judgment, and response reporting (Krosnick, 1991, 1999). Consequently, we assumed that respondents provide shorter and less detailed answers to an open-ended attitude question if they retrieve incomplete information or if they do not put sufficient effort into generating a well-formulated response. In line with previous research, we also suggest that short response latencies might indicate shortcuts and simplifications in the response process (see e.g., Greszki, Meyer, & Schoen, 2015; Roßmann, 2017; Roßmann, Gummer, & Silber, 2018; Smyth et al., 2009).

If respondents pursue strong satisficing as a response strategy, they completely skip the cognitive processes of question comprehension, information retrieval, and judgment (Krosnick, 1991, 1999). Therefore, shortcutting the processes of comprehension can result in answers that do not correspond to the question. Furthermore, a failure to retrieve information and integrate it into a judgment may tempt respondents to provide a response that lacks interpretability because it contains

non-substantive information, such as "don't know," or nonsense entries, such as, for example, ":-)" (cf. Baker et al., 2010; Revilla & Ochoa, 2015b).

Measuring the Response Quality of Open-Ended Attitude Questions

Prior research on open-ended questions in self-administered surveys has used several indicators to study response quality. However, in most instances, these indicators were not derived from a unifying theoretical framework that links respondents' response strategies with the quality of their responses.

First, previous studies often have related the accuracy of a response to its extensiveness, reasoning that the longer an open-ended response is, the more detailed and informative (e.g., Galesic, 2006; Galesic & Bosnjak, 2009; Grauenhorst, Blohm & Koch, 2016; Kwak & Radler, 2002; Mavletova, 2013; Rada & Dominguez-Alvarez, 2013). However, we need to be aware that longer responses are not necessarily more accurate than shorter ones. What we consider to be a high quality response also depends on the type of open-ended question. This study investigated open-ended attitude questions, specifically those that asked about the most important problem facing a country. For this particular type of open-ended question, shorter answers may be sufficient for accurately expressing an attitude, compared to open-ended questions that ask for more narrative responses. Depending on the content of a question, an inherent trade-off may exist between the extensiveness and accuracy of a response: Up to a certain point, the accuracy of a response increases with its length. However, at some point, a further increase in length may indicate that respondents put insufficient effort into integrating their retrieved information into a summarizing judgment. In these cases, it is often difficult to identify the information that the question asked for. Thus, response length alone may not be an ideal indicator of response quality.

Second, a growing body of research has suggested that longer response latencies indicate more effortful and thorough cognitive processing and, as a consequence, a higher quality response, compared to shorter response times (see, e.g., Greszki et al., 2015; Roßmann, 2017; Roßmann et al., 2018; Smyth et al., 2009). Accordingly, existing research has contended that respondents who do not put much effort into answering an open-ended question will tend to write less and have shorter response times (e.g., Revilla & Ochoa, 2015b). However, we have to acknowledge that longer latencies also may signal response problems or flawed questions (Bassili & Scott, 1996). For instance, some respondents might have difficulties understanding and answering a question because its wording is not concise or because it addresses several different topics at once. In this case, longer response latencies do not necessarily indicate higher quality. Short response latencies may also be the result of highly accessible attitudes and, thus, indicate responses of high

quality (cf. Fazio, 1990; Mayerl, Sellke, & Urban, 2005). Thus, we suggest that response latencies should not be used as the sole indicator of response quality.

Third, only a few studies have examined the richness of detail and interpretability of responses to open-ended questions. Two of these studies looked at non-substantive and nonsense answers (Mavletova, 2013; Revilla & Ochoa, 2015a), and Smyth et al. (2009) coded the content of open-ended answers with regard to the number of themes that respondents addressed and the additional elaboration they provided. For this purpose, Smyth et al. (2009) defined a *theme* as "a concept or subject that answered the question and was independent of all other concepts within the response" (p. 327). In line with their reasoning, we suggest that answers that cannot be interpreted (i.e., answers that do not constitute a theme) indicate low quality, since they lack informative content or do not correspond to the question at all.

Although response length and, particularly, response latency are essentially process-generated measures, the interpretability of answers is a content-related indicator of response quality. In our view, this distinction is important because the different indicators of response quality may convey different information, and thus, their use in analyses might lead to different or even contradictory conclusions. As we assume throughout the present study, the interpretability of answers is likely a more appropriate indicator of response quality, compared to response length or latency, because it is less sensitive to conflicting assumptions about its association with quality. However, with regard to the majority of previous research and the naive expectations derived from satisficing theory, we base our analyses on the assumption that longer answers and longer response times reflect higher response quality.

Effects of Survey and Respondent Level Characteristics on the Quality of the Answers to an Open-Ended Attitude Question

In this section, we draw on *satisficing* as a theoretical framework to derive a comprehensive set of hypotheses to address the effects of explanatory factors on both survey and respondent levels (for an overview of our hypotheses, see Table 1).

Survey Level Characteristics

With respect to the survey level, an important factor is the location of the open-ended question in the questionnaire. According to satisficing theory, the response burden accumulates over the course of a questionnaire, which in turn may lead to respondent fatigue (Krosnick, 1991, 1999). Consequentially, the later an open-ended question is placed in a questionnaire, the higher are the chances that respon-

Table 1 Overview of the hypotheses

	Hypothesis
Survey level	
Hypothesis 1	The later open-ended questions are asked in a survey, the lower will be the response quality.
Hypothesis 2	The closer a survey is conducted to an event that is related to a question topic, the higher will be the response quality.
Hypothesis 3	The respondents of a probability-based online panel provide better quality answers, compared to the respondents of an opt-in online panel.
Respondent level	
Hypothesis 4	Higher educated respondents give better quality answers to open-ended questions, compared to less educated respondents.
Hypothesis 5	Older respondents give lower quality answers to open-ended questions, compared to younger respondents.
Hypothesis 6	Highly motivated respondents give better quality answers, compared to less motivated respondents.
Hypothesis 7	Respondents using a mobile device to answer open-ended survey questions give lower quality answers, compared to respondents using a PC.
Cross-level interactions	
Hypothesis 8	The later open-ended questions are asked in a survey, the larger is the effect of the respondents' motivation on response quality.
Hypothesis 9	The closer a survey is conducted to an event that is related to a question topic, the smaller is the effect of the respondents' abilities on response quality.

dents already will be fatigued and that they will perceive answering the question as taxing. Thus, we expect that the later open-ended questions are asked in a survey, the lower will be the response quality (Hypothesis 1).

Another factor at the survey level is the context of an interview. Surveys are conducted within broader societal environments that are characterized by events of which at least some will receive significant attention by the population under study. According to satisficing theory, it can be expected that if relevant information or pre-formulated attitudes are easily accessible, respondents should be motivated to optimize their responses – specifically, their cognitive processes of information retrieval should require much less effort (Krosnick, 1991, 1999). If a topic-related event occurs in close proximity to a survey, respondents should have more easily accessible information. Thus, we hypothesize that the closer a survey is conducted

to an event that is related to a question topic, the higher will be the response quality (Hypothesis 2).

Further, we assume that opt-in online panelists and probability-based panelists differ in the quality of their responses. For example, a study by Silber, Lischewski, and Leibold (2013) compared the response behavior of the professional respondents of two online access panels with the less professional respondents of two web surveys. Their results showed that the respondents of the online access panels had lower break-off rates and were more likely to answer an open-ended attitude question. However, their answers were shorter and less often meaningful compared to the responses of the less professional respondents (Silber et al., 2013). In addition, due to the self-selection in the recruitment process, members of the opt-in online panels were more likely to hold multiple memberships in different online panels (Hillygus, Jackson, & Young, 2014). Therefore, we assume that the respondents from the opt-in online panels are used to answering large quantities of surveys. Moreover, since opt-in panelists presumably do more web surveys, compared to probability-based panelists, they may be less motivated to work through all four steps of cognitive processing, and satisfice more often (Baker et al., 2010). Thus, we expect that the respondents of a probability-based online panel to provide answers of better quality to open-ended questions, compared to the respondents of an opt-in online panel (Hypothesis 3).

Respondent Level Characteristics

With respect to the respondent level, we expect a set of individual characteristics to affect the efforts of respondents to form and report an interpretable response. According to our theoretical framework, respondents with greater ability are used to performing complex mental processes; they are practiced at thinking about the topic of a question and in formulating judgments (Krosnick, 1991, 1999). Previous research has shown that older respondents and those with lower levels of education often provide answers of worse quality (Couper & Kreuter, 2013; Denscombe, 2007; Knäuper, 1999; Loosveldt & Beullens, 2013; Olson & Peytchev, 2007; Roßmann et al., 2018; Yan & Tourangeau, 2008). Thus, we assume that higher educated and younger respondents give answers of better quality to open-ended questions, compared to less educated and older respondents, respectively (Hypothesis 4 and Hypothesis 5).

At the respondent level, another important factor is a respondent's motivation to answer questions accurately. Motivated respondents are more likely to perform all steps of the response process thoroughly, and thus, take their time to read and answer open-ended questions. In line with this assumption, previous studies have suggested that less motivated respondents give faster and shorter responses (e.g., Galesic & Bosnjak, 2009; Holland & Christian, 2009; Olson & Peytchev, 2007).

Thus, we hypothesize that highly motivated respondents give answers of higher quality, compared to less motivated respondents (Hypothesis 6).

In the past decade, the usage of Internet-capable mobile devices, like smartphones and tablets, has increased substantially (Gummer, Quoß, & Roßmann, 2019). Previous research has demonstrated that the use of these mobile devices affects response quality (De Bruijne & Wijnant, 2013; Mavletova, 2013; Peytchev & Hill, 2010; Stapleton, 2013). In particular, with regard to screen size, smartphones differ considerably from personal computers (PCs) and tablets. Since a smaller screen size limits the amount of visible information, respondents sometimes need to scroll or zoom to see the whole question. In addition, selecting a response on a touch screen may take longer due to the smaller screen size (Couper & Peterson, 2017). Thus, answering survey questions on a smartphone may require more effort from respondents (Couper & Peterson, 2017; De Bruijne & Wijnant, 2013; Mavletova, 2013; Peytchev & Hill, 2010; Stapleton, 2013) and therefore increase the burden of providing open-ended responses. Apart from that, respondents may use their smartphones and tablets more often to respond to surveys when they are outside of their home (Mavletova, 2013), and they may be more likely to multitask while completing web surveys (Couper & Peterson, 2017). Therefore, distractions or interruptions may be more common among users of mobile devices, which in turn can negatively affect response quality. This scenario is particularly important with respect to open-ended questions because users of smartphones or tablets usually need to enter their answer on a virtual keyboard, which often is more difficult, and thus, slower than using a regular keyboard with a desktop or notebook computer. In line with these assumptions, studies by Mavletova (2013) and Lugtig and Toepoel (2016) found that the use of smartphones to answer web surveys was associated with shorter responses to open-ended questions. Thus, we expect respondents using a mobile device to give answers of lower quality to open-ended attitude questions, compared to respondents using a PC (Hypothesis 7).

Cross-Level Interactions

Although the factors discussed above are conceptually located at different levels, we assume that they interact. According to satisficing theory, respondents differ in their response strategy depending on the position of the open-ended questions in the survey, and their motivation (Krosnick, 1991, 1999). Whereas higher motivated respondents probably invest more effort in answering open-ended questions, regardless of their position in the survey, less motivated respondents are likely to experience respondent fatigue earlier and switch their response strategy to satisficing (cf. Hypothesis 6). The closer an open-ended question is located near the end of the questionnaire, the larger are the differences between the respondents who are low in motivation and those who are highly motivated. Thus, we assume that the

later the open-ended questions are asked in a survey, the larger is the effect of the respondents' motivation on response quality (Hypothesis 8).

Similarly, we expect an interaction between the proximity of a survey to a topically relevant event and respondents' ability to answer thoroughly an open-ended question on that topic. We also assume that the increased availability of topic-related attitudes and information diminishes the differences between highly able and less able respondents. In this regard, we hypothesize that the closer a survey is conducted to an event that is related to a question topic, the smaller is the effect of the respondents' abilities on response quality (Hypothesis 9).

Data

The present study draws on pooled data from 29 cross-sectional web surveys that were conducted between 2009 and 2015 as part of the German Longitudinal Election Study (Rattinger et al., 2009-2015). Building on the foundations of a repeated cross-section design, key questions were asked repeatedly in each survey, which covered topics such as political attitudes and behaviors, and socio-demographics. Surveys 1-16 used samples from a large German opt-in online panel with about 65,000 to 100,000 active panelists who were recruited to answer surveys on specific issues via online advertisements or via blogs and social media channels. In contrast, surveys 17-29 were sampled from a German probability-based online panel that was comprised of about 40,000 active panelists who were recruited at the end of regular computer-assisted telephone surveys (CATI) that drew on random digit dialing sampling. Comparable quotas on age, sex, and education were used to select each of the 29 samples for the web-based cross-sectional surveys. Accordingly, we calculated each survey's completion rate (AAPOR, 2016) following the recommendations of Callegaro and DiSogra (2008). On average, the completion rate was 82% (for details, see Appendix Table A.1). The pooled data set had 32,494 respondents (~1,120 per survey).

For our analyses, we selected a question measuring public opinion that is regularly asked in open-ended form in surveys (cf. Schuman & Presser, 1996): a question about the most important problem facing the country. The wording of the question was the same for all 29 surveys: "In your opinion, what is the most important political problem facing Germany at the moment?" The original German wording was: "Was ist ihrer Meinung nach gegenwärtig das wichtigste politische Problem in Deutschland?" While the wording of the question was constant across the surveys, the design of the question was slightly changed in some surveys. From survey 18 onwards, the maximum length of respondents' answers was technically limited to 100 characters, which forced respondents to shorten their response. Also, in surveys 21-24, the question was supplemented with additional features that made

respondents aware of the response length limit. In our analyses, we included these changes in design as controls (see below). Answering the open-ended question was voluntary in each of the 29 surveys, so respondents could decide whether they would give an answer or leave the text box empty.

We created three indicators of the quality of the responses to the open-ended attitude question that served as dependent variables in our analyses: response length, response latency, and the interpretability of the answers. We operationalized response length by counting the number of characters. Since the character-based measure of length was skewed to the right (*Skewness*=2.64), we used the natural logarithm of the length for our further analyses. This transformation reduced the skewness to 0.44.

We measured response latency to the open-ended question in seconds.[1] As before, we used the natural logarithm to account for the skewness of the response latency measure. This reduced the skewness from 1.40 to -0.65.

Furthermore, we used the interpretability of the responses as a content-related indicator of response quality. During data processing, we coded respondents' answers to the open-ended questions into categories using a predefined coding scheme developed and extensively tested by the project team of the German Longitudinal Election Study.[2] We used the categories of this coding scheme to create a dummy variable that indicated whether the answers were interpretable or not (0 = not interpretable / 1 = interpretable). Answers that could not be interpreted (e.g., "asdf", "---"), did not mention a problem ("don't know"), or represented a refusal were coded as not interpretable. Answers that corresponded to the question and mentioned specific themes (e.g., "unemployment") were coded as substantive responses.

To explain response quality, we drew on a set of independent variables at the survey and respondent level. In addition, we included two cross-level interactions. Table 2 presents the descriptive statistics for all the variables we used in our analy-

1 An issue with response latencies is that their distributions are almost inevitably skewed (Fazio, 1990). Particularly in the absence of an interviewer in web-based surveys, we observed a characteristic long tail of slow latencies in the distribution. Since we do not know whether extremely slow latencies are caused by situational factors (e.g., distractions) or by lower abilities of respondents, we used a common outlier detection method and, in each survey, set response latencies that were longer than the mean plus two times the standard deviation of the distribution to missing (see e.g., Bassili & Fletcher, 1991). Therefore, we first omitted extreme outliers (5 minutes or more to answer the question) that would have skewed the distribution and affected the mean-based outlier criterion. Applying this approach, we classified 7.9% of the data points as response time outliers.

2 The development of the coding scheme for the open-ended question on the most important problem facing Germany was complemented by extensive tests of inter-coder reliability. Then, the coders received a comprehensive coding scheme, which included further information and detailed coding instructions to ensure high coding quality.

ses. A more detailed discussion on the operationalization of respondent level variables is provided in Appendix B.

At the survey level, we included three variables to test Hypotheses 1, 2, and 3. First, the web surveys used in this study applied a paging design (Couper, 2008). Thus, the screen on which the open-ended question appeared was a very good estimate of its position in the questionnaire (Hypothesis 1). For instance, in survey 29, the open-ended question appeared on the 10th screen. For an easier interpretation of the effects in our models (see Section 5), we rescaled the variable to a range of 0 to 1. Second, we included a dummy variable that indicated whether a survey was conducted within 6 months before or after the German federal elections in the years 2009 and 2013 (Hypothesis 2). Elections are among the most important political events in democratic societies. Since both the open-ended attitude question and web surveys were strongly related to political issues and elections in particular, it is likely that respondents have more readily available attitudes in times when a multitude of these issues are central to the public debate. Political information should be highly available for respondents due to election campaigns, which they can follow on advertising posters, television, or the Internet. In addition, during election campaigns, the appearance of specific political issues in the media and their handling by the candidates is higher (Huber, Rattinger, & Wagner, 2009; Schumann & Schoen, 2009). Third, we created a dummy variable that indicated whether the survey used respondents from an opt-in (surveys 1–16) or a probability-based (surveys 17–29) online panel (Hypothesis 3). For controls, we included two variables that indicated whether the response length was technically limited (0 = no / 1 = yes) and whether the question was supplemented with additional features to make respondents aware of the 100 characters response length limit (0 = no / 1 = yes).

With respect to the respondent level, we used education (0 = low / 1 = intermediate / 2 = high) as an indicator of the respondent's ability (Hypotheses 4). Since we also assumed that ability is associated with age, we included it (0 = 18–29 / 1 = 30–39 / 2 = 40–49 / 3 = 50–59 / 4 = 60+) as a second indicator (Hypothesis 5). Hypothesis 6 suggests that a respondent's motivation may influence their response behavior. Accordingly, we included three related variables: interest in the survey topic (0 = low interest in politics / 1 = intermediate interest in politics / 2 = high interest in politics), strength of the respondent's identification with a political party (0 = none / 1 = moderate / 2 = strong), and (intended) turnout to vote in a federal election (0 = no / 1 = yes). To examine the effects of different devices on response quality (Hypothesis 7), we identified whether respondents used a PC (desktop or notebook), tablet, or smartphone to complete the survey. The information on the device was extracted from the user agent string using the Stata command parseuas (Roßmann & Gummer, 2016). For control variables, we included the respondent's sex (0 = male / 1 = female) and region of residence (0 = East Germany / 1 = West Germany).

Table 2 Variables used to explain the response quality of open-ended attitude questions

Variable	M	Min	Max	N
Survey level				
Position of open-ended question	0.15	0	1	29
Proximity to election	0.41	0	1	29
Probability-based online panel	0.45	0	1	29
Respondent level				
Age				
18–29	0.23	0	1	32,494
30–39	0.20	0	1	32,494
40–49	0.24	0	1	32,494
50–59	0.16	0	1	32,494
60+	0.16	0	1	32,494
Education				
low	0.31	0	1	32,209
intermediate	0.39	0	1	32,209
high	0.30	0	1	32,209
Interest in politics				
low	0.21	0	1	32,458
intermediate	0.40	0	1	32,458
high	0.39	0	1	32,458
Intention to vote	0.85	0	1	32,449
Strength of party identification				
none	0.28	0	1	32,426
moderate	0.28	0	1	32,426
strong	0.44	0	1	32,426
Device				
personal computer	0.93	0	1	32,491
smartphone	0.04	0	1	32,491
tablet	0.03	0	1	32,491
Control Variables				
Technical limit of answer to 100 characters	0.41	0	1	29
Information on 100 characters limit	0.14	0	1	29
Sex: Female	0.50	0	1	32,494
Region: West Germany	0.80	0	1	32,487

Note. M = mean. Statistics at the respondent level variables are calculated with N = number of respondents. Statistics at the survey level are calculated with N = number of surveys.

As argued previously in the present study, interactions between respondent and survey characteristics can be assumed to partially explain response behavior. Thus, we created a cross-level interaction between the location of the open-ended attitude question in the questionnaire and the respondents' interest in the survey topic (Hypothesis 8). Further, to test whether topic-related events enhance the availability and accessibility of relevant attitudes and information, we created a second cross-level interaction between the survey's proximity to a federal election and the respondents' ability as indicated by their level of education (Hypothesis 9).

Methods

To statistically account for the multilevel structure of our data – individuals clustered in surveys – and to test the hypotheses and interactions of two conceptual levels (respondent and survey level), we applied multilevel modeling (Hox, 2010; Luke, 2004; Rabe-Hesketh & Skrondal, 2008; Snijders & Bosker, 1999) using Stata 14.1. This approach explicitly modeled that the characteristics of the lower level (i.e., respondents) depend on the higher level (i.e., surveys). Our mathematical expressions mainly refer to the work of Snijders and Bosker (1999) and Luke (2004).

We fitted a random intercept model with fixed slopes and cross-level interactions. Since we assumed that the location of a question in a survey and the proximity of a survey to a topic-related event explain the variation in the coefficients of respondents' ability and motivation (Hypotheses 8 & 9), the slopes were fixed. In the following, Y_{ij} denotes an individual i's response behavior in survey j. X_{pij} is a vector of p characteristics at the respondent level, whereas Z_{qj} is a vector of q characteristics at the survey level. $X_{pij}Z_{qj}$ is a vector of cross-level interactions. Thus, γ_{0p}, γ_{q0}, and γ_{qp} are the respective regression coefficients. γ_{00} is the grand mean, u_{0j} is the survey level residuals, and r_{ij} is the respondent level residuals. Consequently, our final (linear) model used to explain response length and latency is denoted in single-equation form as follows:

$$Y_{ij} = \gamma_{00} + \sum \gamma_{0p} X_{pij} + \sum \gamma_{q0} Z_{qj} + \sum \gamma_{qp} X_{pij} Z_{qj} + u_{0j} + r_{ij}$$

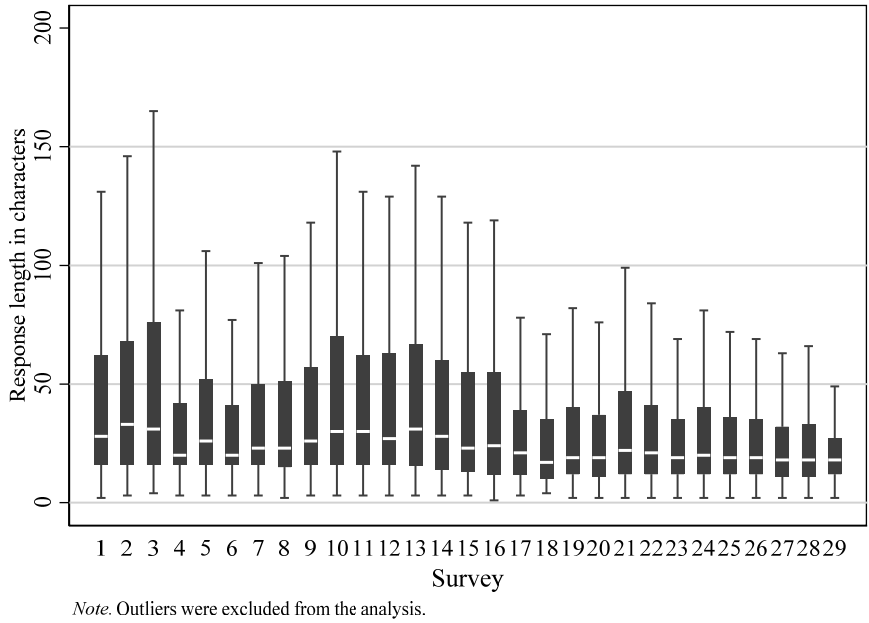
in which $Y_{ij} = \ln\left(Y_{ij}\right)$ is the transformed response length and latency that we used as dependent variables. Due to their operationalization, we assumed that both response length and latency indicators are approximately normally distributed.

Since our theoretical reasoning remained the same for our binary dependent variable, the respective logistic multilevel model also was specified as a random intercept fixed slope model with cross-level interactions. Accordingly, we modelled the probability of respondent i giving an interpretable answer P_{ij} in survey j as follows:

$$logit(P_{ij}) = \gamma_{00} + \sum \gamma_{0p} X_{pij} + \sum \gamma_{q0} Z_{qj} + \sum \gamma_{qp} X_{pij} Z_{qj} + u_{0j}$$

Results

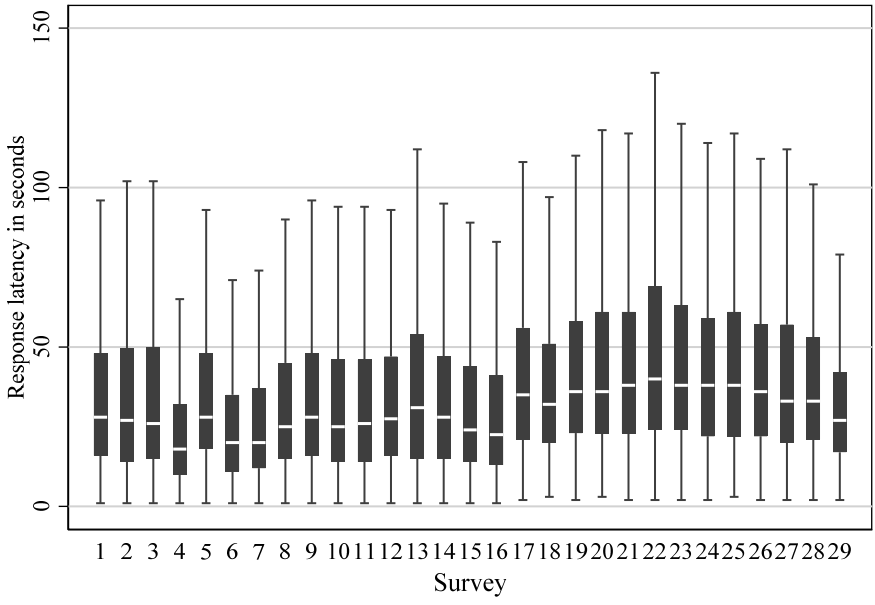
Before reporting the results of our multilevel models, we will briefly discuss the variation in the length of responses, response latencies, and interpretable answers across surveys. Figure 1 shows the length of the responses and visualizes the variations between surveys, which is particularly evident for the surveys 18-29 in which the response length was limited to 100 characters.



Note. Outliers were excluded from the analysis.

Figure 1 Boxplots of the response length to the question on the most important political problem facing Germany for 29 web surveys

The boxplots of Figure 2 illustrate the variations in response latencies across surveys. These plots suggest that a strong variation exists in latencies within each survey, and between surveys. Apparently, the average response latencies increased after the sampling switched from an opt-in panel to a probability-based panel in survey 17.



Note. Outliers were excluded from the analysis.

Figure 2 Boxplots of response latency to the question on the most important political problem facing Germany for 29 web surveys

Figure 3 depicts the variation in the share of interpretable answers to the open-ended question. On average, 88.05% of the answers were interpretable with a strong variation across surveys. Notably, more between-survey variation of interpretable answers seems to occur when an opt-in panel was used; and a more homogeneous (but larger) share of interpretable answers seems to occur between surveys when a probability-based panel was used. Table 3 shows the results of the multilevel models for the three indicators of response quality: response length (Model 1), response latency (Model 2), and interpretability of the answers (Model 3).

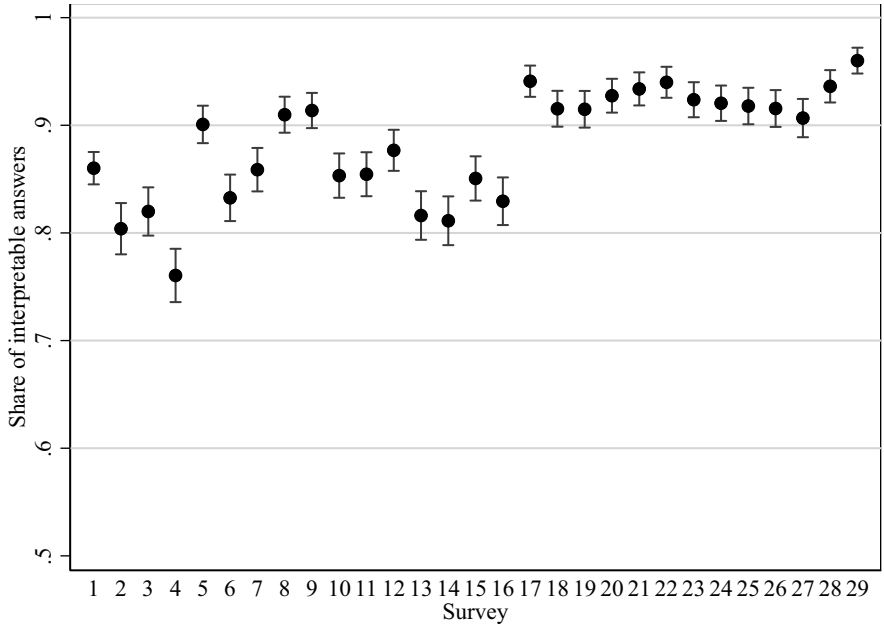


Figure 3 Share of interpretable answers to the question on the most important political problem facing Germany for 29 web surveys

Note that all variables reported in Table 3 range from 0 to 1. Thus, the coefficients of Model 1 and 2 provide the marginal effects of each of the dependent variables. This enables an easy interpretation of the coefficients, since a 1-unit change in any of the independent variables is equivalent to comparing a respondent at the minimum value of the respective variable to a respondent at the maximum value. Similarly, we report the average marginal effects (AMEs) for logistic Model 3. AMEs enable an intuitive interpretation as the average effect on the probability over all cases in the sample (Best & Wolf, 2015).[3]

3 We tested multiple hypotheses in our study what could possibly result in capitalizing on chance (Type I error, i.e., rejecting too many null hypotheses). However, popular adjustment methods such as the Bonferroni correction come at the price of lowering statistical power and, in turn, increase the chance of Type II errors (Gelman, Hill, & Yajima, 2012; Rothman, 1990). Against this caveat, we remain skeptical whether to correct for this potential issue. In addition, we derived all hypotheses from satisficing theory and prior research, what lays ground for a careful assessment of the plausibility of effects that we found to be statistically significant. In our view, this approach should limit the negative consequences that potential Type I errors might have. Accordingly, we argue that the problem of multiple comparisons is not likely to be a major issue in the present study.

To assess the explanatory power of our models, we calculated the intraclass correlation coefficient (ICC) for all three dependent variables based on empty models. For the response length (ICC = .05), the response latency (ICC = .07), and the interpretability of the answers (ICC = .07), part of the variance can be attributed to the survey level. The residual variances of our three final models (Table 3) further indicate that including our covariates reduced the proportion of unexplained variation between surveys.

Table 3 Multilevel models of the response quality to open-ended attitude questions

	Model 1 Response Length	Model 2 Response Latency	Model 3 Response Interpretability	
	b (SE)	b (SE)	b (SE)	AME (SE)
Survey level effects				
Position of open-ended question	-0.041	-0.452***	-0.886***	-0.083***
	(0.104)	(0.107)	(0.241)	(0.023)
Proximity to election	0.034	0.123*	0.466***	0.044***
	(0.048)	(0.053)	(0.125)	(0.012)
Probability-based online panel	-0.296***	0.343***	0.918***	0.086***
	(0.089)	(0.100)	(0.244)	(0.023)
Respondent level effects				
Education: low	Ref.	Ref.	Ref.	Ref.
intermediate	0.041*	-0.006	0.259***	0.024***
	(0.018)	(0.016)	(0.059)	(0.006)
high	0.111***	0.049**	0.463***	0.044***
	(0.019)	(0.017)	(0.069)	(0.007)
Age: 18–29	Ref.	Ref.	Ref.	Ref.
30–39	-0.026	0.079***	0.262***	0.025***
	(0.017)	(0.015)	(0.050)	(0.005)
40–49	-0.043**	0.187***	0.662***	0.062***
	(0.016)	(0.014)	(0.052)	(0.005)
50–59	-0.050**	0.247***	0.817***	0.077***
	(0.018)	(0.016)	(0.064)	(0.006)
60+	0.056**	0.405***	0.896***	0.084***
	(0.018)	(0.017)	(0.070)	(0.007)
Interest in politics: low	Ref.	Ref.	Ref.	Ref.
intermediate	0.102***	0.177***	0.681***	0.064***
	(0.018)	(0.016)	(0.052)	(0.005)
high	0.210***	0.228***	1.354***	0.127***
	(0.019)	(0.017)	(0.069)	(0.007)

	Model 1 Response Length	Model 2 Response Latency	Model 3 Response Interpretability	
	b (SE)	b (SE)	b (SE)	AME (SE)
Intention to vote	-0.065*** (0.017)	0.039** (0.015)	0.233*** (0.046)	0.022*** (0.004)
Strength of party identification: none	Ref.	Ref.	Ref.	Ref.
moderate	-0.048** (0.015)	0.083*** (0.013)	0.387*** (0.046)	0.036*** (0.004)
strong	-0.082*** (0.014)	0.082*** (0.013)	0.657*** (0.047)	0.062*** (0.005)
Device: personal computer	Ref.	Ref.	Ref.	Ref.
smartphone	-0.163*** (0.028)	0.129*** (0.025)	0.073 (0.102)	0.007 (0.010)
tablet	-0.152*** (0.031)	0.010 (0.029)	0.110 (0.135)	0.010 (0.013)
Cross-level interaction effects				
Interest in politics: intermediate ×	0.022	0.056	-0.361*	-0.034*
Position of open-ended question	(0.067)	(0.054)	(0.150)	(0.014)
Interest in politics: high ×	-0.009	0.122*	-0.510**	-0.048**
Position of open-ended question	(0.066)	(0.054)	(0.170)	(0.016)
Education: intermediate ×	-0.056*	-0.023	-0.076	-0.007
Proximity to election	(0.026)	(0.023)	(0.084)	(0.008)
Education: high ×	-0.053	-0.076**	-0.247*	-0.023*
Proximity to election	(0.028)	(0.025)	(0.098)	(0.009)
Control variables				
Technical limit of answer to 100 characters	-0.073 (0.091)	-0.039 (0.102)	-0.206 (0.248)	-0.019 (0.023)
Information on 100 characters limit	0.091 (0.059)	0.059 (0.066)	-0.167 (0.155)	-0.016 (0.015)
Sex: female	0.026* (0.011)	-0.046*** (0.010)	-0.071 (0.038)	-0.007 (0.004)
Region: West Germany	-0.007 (0.013)	-0.046*** (0.012)	-0.157** (0.048)	-0.015** (0.005)
Intercept	3.346*** (0.039)	2.820*** (0.039)	0.132 (0.102)	
$\sigma^2_{u_0}$	0.006	0.008	0.036	
σ^2_r	0.781	0.683	3.290	
N	28,264	29,520	32,062	

Note. *p*-values: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Survey level effects

With respect to the location of the open-ended question in the questionnaire (Hypothesis 1), we did not find any evidence that later placement affected the length of answers. However, respondents took less time to answer the question the later it was placed in a questionnaire. Further, respondents gave interpretable answers to a significantly lesser extent the later the open-ended question was asked in the survey. The latter two findings are in line with the expectation that fatigue increases over the course of a questionnaire, which increases the likelihood of respondents adopting a satisficing response strategy.

Also, we hypothesized that if the survey was conducted in proximity to a topic-related event (i.e., the German federal elections 2009 and 2013), respondents would provide answers of higher quality due to the increased availability and accessibility of pre-formulated attitudes and relevant information (Hypothesis 2). The results of Models 2 and 3 showed that respondents took more time to answer and that responses were more likely to be interpretable if the survey was carried out 6 months before or after an election, which we hypothesized to be an effect of more accessible information. We did not observe significant effects on response length.

Next, we expected that respondents of a probability-based online panel would provide higher quality responses to an open-ended attitude question, compared to opt-in panelists (Hypothesis 3). Again, our findings are mixed. We found that membership in the probability-based online panel had a negative effect on response length. However, this negative effect was not particularly surprising, since the introduction of the 100 character limit in survey 18 almost perfectly coincided with the change of the panel provider in survey 17. Accordingly, we refrain from overinterpreting this finding. In contrast, we found a positive effect of membership in a probability-based online panel on response latency. Our results also revealed that respondents of the probability-based online panel gave interpretable answers at a significantly higher rate than the opt-in panelists. The latter findings supported our theoretical expectation that the sample of the probability-based online panel was composed of less over-surveyed, and thus, more motivated respondents who engaged in providing interpretable responses, compared to the sample of the opt-in online panel.

Respondent level effects

As we had expected in Hypothesis 4, our results confirmed that highly educated respondents gave answers of higher quality. On average, their answers were longer, and they took more time to respond to the open-ended question, compared to less educated respondents. Higher educated respondents also gave interpretable answers at a higher rate. These findings are in line with the assumption that respondents high in ability are more likely to carefully execute all steps of cognitive processing. In contradiction to Hypothesis 5, we found that the group of the

oldest respondents needed more time to answer the open-ended question and gave interpretable responses to a greater extent, compared to younger age groups. On the basis of these findings, we rejected Hypothesis 5. On the one hand, the lower response quality of younger respondents was surprising, since previous studies (e.g., Knäuper, 1999) and satisficing theory (Krosnick, 1991) have suggested that younger respondents tend to provide better responses due to a higher working memory capacity (i.e., ability). On the other hand, we interpret our results as an indication that age might not be a well-suited measure for determining respondents' abilities to thoroughly answer open-ended questions (cf. Holbrook, Krosnick, Moore, & Tourangeau, 2007).

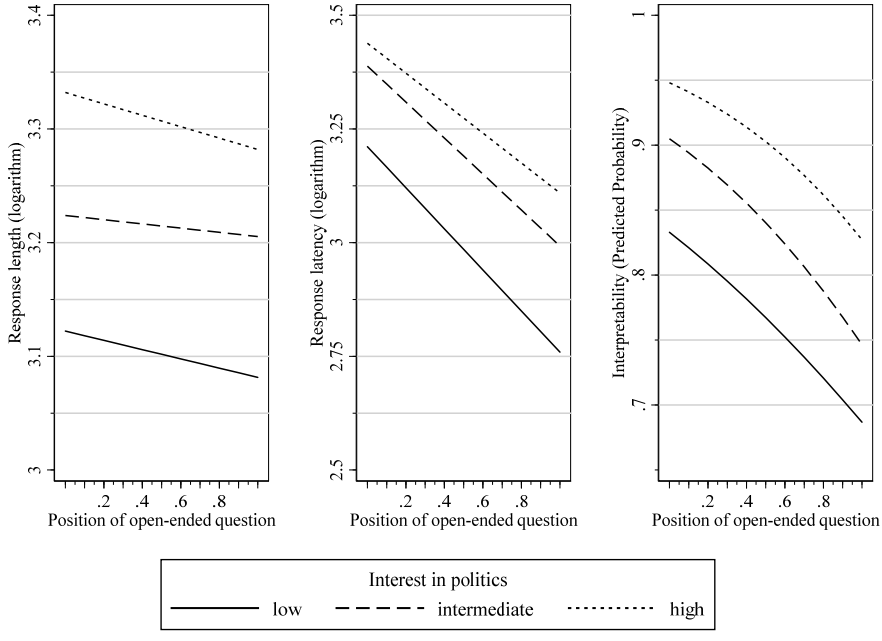
With regard to motivation (Hypothesis 6), we found that respondents with a higher interest in a survey topic provided a significantly better response quality than less interested respondents. They gave longer answers, took more time to respond, and provided interpretable answers at a higher rate. These results indicate that interest in a survey topic plays an important role in shaping the quality of the responses to open-ended questions (see, e.g., Holland & Christian, 2009). In contradiction to our expectations, respondents who turned out, or intended to vote gave significantly shorter responses, compared to those who did not intend to vote. A similar pattern emerged with respect to respondents' identification with a political party: respondents who reported a strong or at least a moderate psychological attachment gave significantly shorter answers than those who did not identify themselves with a party at all. However, in line with Hypothesis 6, a moderate or strong party identification had significant positive effects on the response latency and interpretability of the response. Thus, our findings regarding the effects of motivation on response quality largely confirm Hypothesis 6.

In line with previous studies (e.g., Lugtig & Toepoel, 2016; Mavletova, 2013), our results showed that tablet or smartphone usage negatively affected the number of characters entered, compared to the use of a PC (desktop or notebook). Smartphone users also took more time to answer the open-ended attitude question than respondents using a PC. As discussed previously in the present study, respondents may take longer answering survey questions with a smartphone due to the smaller screen size and the use of virtual keyboards (e.g., Couper & Peterson, 2017; Mavletova, 2013). However, we found no significant effects of mobile device usage on the rate of interpretable answers. Thus, Hypothesis 7 was only partly confirmed with respect to the length of answers and response latency for smartphone users. We suggest that these findings indicate that the use of mobile devices - particularly smartphones - has notable effects on the process of entering open-ended responses, but not necessarily on the quality of the content.

Cross-level interaction effects

In the last step of our analyses, we examined whether the respondent and survey level factors interacted across conceptual levels to affect the quality of the responses to open-ended questions.

In particular, we assumed that the later the open-ended attitude questions are asked in a survey, the larger is the effect of the respondents' motivation on their response quality (Hypothesis 8). For the purpose of illustration, Figure 4 presents interaction plots for each of the three indicators.

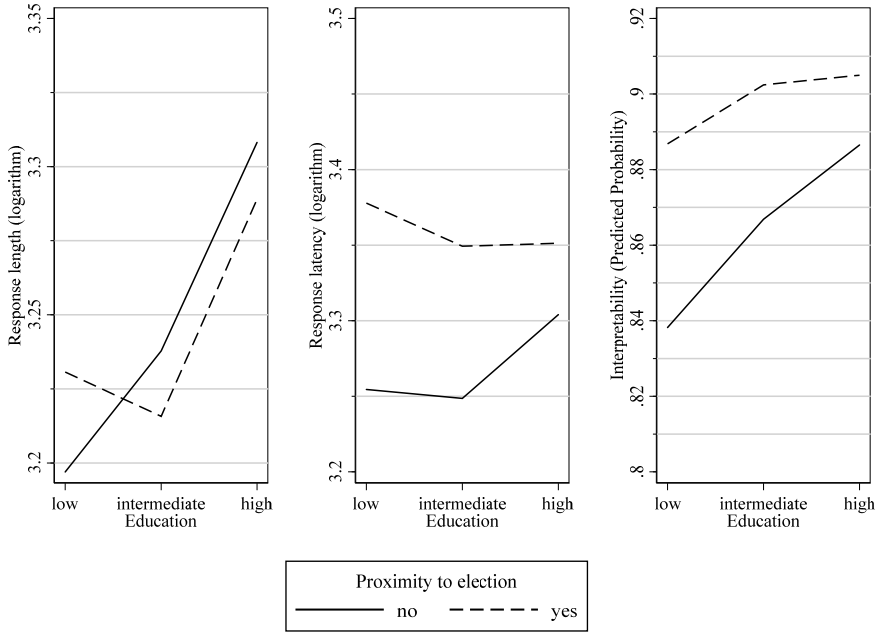


Note. Predicted values based on models presented in Table 3: Model 1 = response length, Model 2 = response latency, and Model 3 = interpretability of response.

Figure 4 Cross-level interactions between the position of an open-ended question in a survey and the effects of respondents' motivation on response quality

In contrast to our expectation, we found no significant effects on response length due to the cross-level interaction of respondents' interest in a survey topic and the location of the open-ended attitude question in the questionnaire (Figure 4, first plot). Although Models 2 and 3 showed significant interaction effects on the response latency (for high political interest) and interpretability of responses, the visual presentation of these effects (Figure 4, second and third plot) suggests that

the differences between respondents who are low in motivation and those who are highly motivated is only slightly different if the open-ended attitude question is located towards the end of the questionnaire. Consequently, even though we found that effects occurred, their impact seems to be limited. With regard to Hypothesis 9, we found a significant negative effect on the length of answers due to the interaction between the intermediate levels of education and the proximity of a survey to a federal election (Figure 5, first plot).



Note. Predicted values based on models presented in Table 3: Model 1 = response length, Model 2 = response latency, and Model 3 = interpretability of response.

Figure 5 Cross-level interactions between the proximity of a survey to a federal election and respondents' education level

In addition, significant impacts occurred on the interpretability of responses and response latencies due to the interaction of high education and the proximity of a survey to a federal election (Figure 5, second and third plot). With respect to the respondent level, the study found that highly educated respondents provided answers of higher quality, compared to lower educated respondents (cf. Hypothesis 3), and all respondents gave more interpretable answers during a time period of 6 months before or after a federal election (cf. Hypothesis 2). However, the cross-level interaction effects imply a more complex association. In line with our theoretical expectations, it seemed that in times where political issues were central in the pub-

lic debate, related attitudes and information also were more readily accessible for less educated respondents. The results indicated that the differences in response quality between low and highly educated respondents were reduced by the occurrence of a topic-related event (i.e., a federal election) (Hypothesis 9), albeit only to a small extent as Figure 5 illustrates.

Conclusion

The present study investigated the effects of respondent and survey characteristics on the response quality of open-ended attitude questions in web surveys, which complements previous research in several ways. First, we analyzed a pooled data set of 29 web surveys on the political attitudes and behaviors of German Internet users. These data not only enabled us to study respondent level effects, but also to gain new insights into the effects of survey design, and the interaction of this design and respondent characteristics on the response quality of an open-ended attitude question.

Second, we used three different indicators of response quality. Nevertheless, the results of our analyses with these three indicators did not provide unambiguous evidence for every hypothesis. Thus, the question arises as to whether short responses or latencies consistently imply bad response quality or not. In other words, we need to ask whether the relationship between these indicators and response quality is more complex than the majority of previous studies have suggested (see Section 2). Our results indicate that analyzing the response quality to an open-ended attitude question exclusively with single indicators, for instance, with response length or latency, may create a misleading picture. Including content-related indicators such as the interpretability of responses provided us with more differentiated insights, compared to the exclusive use of process-generated measures of quality (i.e., response length or latency). Moreover, for the majority of the survey and respondent level variables, their effects on the content-related measure of response quality were in line with the theoretical expectations. We believe this result is an indication that the content-related measure captured what is most generally understood as the response quality of open-ended questions. Thus, in future studies on the quality of responses to open-ended questions in surveys, we recommend using content-related indicators, such as the number of themes that were addressed (Smyth et al., 2009) or the interpretability of answers. For future research, studying a variety of response quality indicators and exploring the empirical and theoretical relationships between them certainly seems worthwhile.

Third, we used satisficing theory to analyze the response quality of open-ended questions. The analyses we carried out lend support to several hypotheses on the effects of respondent and survey level characteristics on response quality, which

we derived from the satisficing framework. In particular, our empirical results support the assumption that motivated respondents and those high in ability provided higher quality responses. These results are in line with previous studies that have found that respondents who are more interested in a survey topic or who are more highly educated are more likely to provide an open-ended response of good quality, compared to less motivated or less able respondents (Denscombe, 2007; Holland & Christian, 2009; Knäuper, 1999; Smyth et al., 2009).

Fourth, by including cross-level interactions in our models, we found that factors on different conceptual levels were not completely independent in affecting response quality. This finding emphasizes the need for further studies on the effects on answer quality caused by the cross-level interactions between respondent and survey level characteristics. Moreover, the finding that significant, albeit small, differences exist with respect to interpretable responses - due to the interaction of the location of the open-ended question in the survey and respondents' low and high in interest in a survey topic - highlights the importance of considering a respondent's motivation when designing web surveys. This finding supports the results from experimental studies that have demonstrated that altering the visual design of a survey can stimulate less motivated respondents to provide responses of better quality (cf. Holland & Christian, 2009; Smyth et al., 2009). For example, Smyth et al. (2009) found that using an introduction that emphasizes the importance of answers to the researchers increased the respondents' elaboration of themes. Also, the results of our study indicate that survey designers should take into account the societal context during the data collection period, since the response quality of an open-ended attitude question can be influenced by topic-related events that occur in proximity to the survey (e.g., a federal election). The present study has shown that the occurrence of such an event can diminish the differences in response quality that normally are caused by the differences in respondents' abilities. This finding is particularly important when analyzing (pooled) longitudinal data sets, which are comprised of interviews that were conducted in close proximity to topic-related events and others that were not. As our findings suggest, measurement errors are not homogenous across surveys; instead, they differ systematically. During analyses, these errors may be mistaken for a substantive change over time or surveys.

The following limitations of the present study should, however, be considered. First, with regard to the political topic of a survey and the particular type of open-ended attitude question (the most important problem), we suggest that follow-up studies should further examine how findings can vary across different survey topics or hold for other types of open-ended questions (e.g., open-ended questions that require more narrative responses). Second, in the present study, we limited the number of survey level characteristics because we decided to pool similar surveys. In light of this limitation, future studies could compile a more diversely designed set of surveys to test more interactions of more factors at the respondent and sur-

vey levels. Compiling a larger collection of surveys should help future studies to arrive at findings that are more robust. Finally, a further interesting opportunity for upcoming research would be to develop additional content-related indicators of data quality to measure how strongly responses correspond to the actual open-ended question, and whether these responses are interpretable.

References

AAPOR. (2016). *Standard Definitions: Final Dispositions of Case Codes and Outcome Rates for Surveys*: The American Association for Public Opinion Research.

Baker, R., Blumberg, S. J., Brick, J. M., Couper, M. P., Courtright, M., Dennis, J. M., . . . Zahs, D. (2010). Research Synthesis: AAPOR Report on Online Panels. *Public Opinion Quarterly, 74*(4), 711-781. doi:10.1093/poq/nfq048

Bassili, J. N., & Fletcher, J. F. (1991). Response-Time Measurement in Survey Research. A Method for CATI and a New Look at Nonattitudes. *Public Opinion Quarterly, 55*(3), 331-346. doi:10.1086/269265

Bassili, J. N., & Scott, B. S. (1996). Response Latency as a Signal to Question Problems in Survey Research. *Public Opinion Quarterly, 60*(3), 390-399. doi:10.1086/297760

Best, H., & Wolf, C. (2015). Logistic Regression. In H. Best & C. Wolf (Eds.), *Regression Analysis and Causal Inference* (pp. 153-171). Los Angeles: Sage.

Callegaro, M., & DiSogra, C. (2008). Computing Response Metrics for Online Panels. *Public Opinion Quarterly, 72*(5), 1008-1032. doi:10.1093/poq/nfn065.

Callegaro, M., Yang, Y., Bhola, D., & Dillman, D. A. (2004). *Response Latency as an Indicator of Optimizing. A Study Comparing Job Applicants and Job Incumbents' Response Time on a Web Survey.* Paper presented at the Proceedings of the RC 33 Sixth International Conference on Social Science Methodology, Amsterdam.

Couper, M. P. (2008). *Designing Effective Web Surveys*. New York: Cambridge University Press.

Couper, M. P., & Kreuter, F. (2013). Using Paradata to Explore Item Level Response Times in Surveys. *Journal of the Royal Statistical Society, 176*(1), 271-286. doi:10.1111/j.1467-985X.2012.01041.x

Couper, M. P., & Peterson, G. J. (2017). Why Do Web Surveys Take Longer on Smartphones? *Social Science Computer Review*, 35(3), 357-377. doi:10.1177/0894439316629932

Couper, M. P., Tourangeau, R., Conrad, F. G., & Zhang, C. (2013). The Design of Grids in Web Surveys. *Social Science Computer Review, 31*(3), 322-345. doi:10.1177/0894439312469865

De Bruijne, M., & Wijnant, A. (2013). Comparing Survey Results Obtained via Mobile Devices and Computers: An Experiment with a Mobile Web Survey on a Heterogeneous Group of Mobile Devices Versus a Computer-Assisted Web Survey. *Social Science Computer Review, 31*(4), 482-504. doi:10.1177/0894439313483976

Denscombe, M. (2007). The Length of Responses to Open-Ended Questions: A Comparison of Online and Paper Questionnaires in Terms of a Mode Effect. *Social Science Computer Review, 26*(3), 359-368. doi:10.1177/0894439307309671

Dillman, D. A., Smyth, J. D., & Christian, L. M. (2014). *Internet, Phone, Mail, and Mixed-Mode Surveys: The Tailored Design Method*: John Wiley & Sons.

Fazio, R. H. (1990). A Practical Guide to the Use of Response Latency in Social Psychological Research. In C. Hendrick & M. S. Clark (Eds.), *Research Methods in Personality and Social Psychology* (pp. 74-97). Thousand Oaks, CA: Sage.

Fuchs, M. (2009). Differences in the Visual Design Language of Paper-and-Pencil Surveys Versus Web Surveys A Field Experimental Study on the Length of Response Fields in Open-Ended Frequency Questions. *Social Science Computer Review, 27*(2), 213-227.

Galesic, M. (2006). Dropouts on the Web: Effects of Interest and Burden Experienced During an Online Survey. *Journal of Official Statistics, 22*(2), 313-328.

Galesic, M., & Bosnjak, M. (2009). Effects of Questionnaire Length on Participation and Indicators of Response Quality in a Web Survey. *Public Opinion Quarterly, 73*(2), 349-360. doi:10.1093/poq/nfp031

Gelman, A., Hill, J., & Yajima, M. (2012). Why We (Usually) Don't Have to Worry About Multiple Comparisons. *Journal of Research on Educational Effectiveness, 5*(2), 189-211.

Grauenhorst, T., Blohm, M., & Koch, A. (2016). Respondent Incentives in a National Face-to-Face Survey Do They Affect Response Quality? *Field Methods*, 28(3), 266-283. doi:10.1177/1525822X15612710

Greszki, R., Meyer, M., & Schoen, H. (2015). Exploring the Effects of Removing "Too Fast" Responses and Respondents from Web Surveys. *Public Opinion Quarterly, 79*(2), 471-503. doi:10.1093/poq/nfu058

Gummer, T., Quoß, F., & Roßmann, J. (2019). Does Increasing Mobile Device Coverage Reduce Heterogeneity in Completing Web Surveys on Smartphones? *Social Science Computer Review*, 37(3), 371-384. doi:10.1177/0894439318766836

Gummer, T., & Roßmann, J. (2015). Explaining Interview Duration in Web Surveys: A Multilevel Approach. *Social Science Computer Review, 33*(2), 217-234. doi:10.1177/0894439314533479

Hillygus, S. D., Jackson, N., & Young, M. (2014). Professional Respondents in Nonprobability Online Panels. In M. Callegaro, R. Baker, J. Bethlehem, A. S. Göritz, J. A. Krosnick, & P. J. Lavrakas (Eds.), *Online Panel Research: A Data Quality Perspective* (pp. 219-237): Chichester: Wiley.

Holbrook, A. L., Anand, S., Johnson, T. P., Cho, Y. I., Shavitt, S., Chávez, N., & Weiner, S. (2014). Response Heaping in Interviewer-Administered Surveys: Is It Really a Form of Satisficing? *Public Opinion Quarterly, 78*(3), 591-633. doi:10.1093/poq/nfu017

Holbrook, A. L., Krosnick, J. A., Moore, D., & Tourangeau, R. (2007). Response Order Effects in Dichotomous Categorical Questions Presented Orally. The Impact of Question and Respondent Attributes. *Public Opinion Quarterly, 71*(3), 325-348. doi:10.1093/poq/nfm024

Holland, J. L., & Christian, L. M. (2009). The Influence of Topic Interest and Interactive Probing on Responses to Open-Ended Questions in Web Surveys. *Social Science Computer Review, 27*(2), 196-212.

Hox, J. (2010). *Multilevel Analysis. Techniques and Applications*. New York: Routledge.

Huber, S., Rattinger, H., & Wagner, C. (2009). No Matter When? Eine Analyse von Feldzeiteffekten. In S. Huber, H. Rattinger, & C. Wagner (Eds.), *Vom Interview zur Analyse* (Vol. 1, pp. 231-258).

Knäuper, B. (1999). The Impact of Age and Education on Response Order Effects in Attitude Measurement. *Public Opinion Quarterly, 63*, 347-370.

Krosnick, J. A. (1991). Response Strategies for Coping with the Cognitive Demands of Attitude Measures in Surveys. *Applied Cognitive Psychology, 5*, 213-236.

Krosnick, J. A. (1999). Survey Research. *Annual Review of Psychology, 50*, 537-567.

Kwak, N., & Radler, B. (2002). A Comparison Between Mail and Web Surveys: Response Pattern, Respondent Profile, and Data Quality. *Journal of Official Statistics, 18*(2), 257-273.

Loosveldt, G., & Beullens, K. (2013). 'How Long Will it Take?' An Analysis of Interview Length in the Fifth Round of the European Social Survey. *Survey Research Methods, 7*(2), 69-78.

Lugtig, P., & Toepoel, V. (2016). The Use of PCs, Smartphones, and Tablets in a Probability-Based Panel Survey: Effects on Survey Measurement Error. *Social Science Computer Review, 34*(1), 78-94. doi:10.1177/0894439315574248

Luke, D. A. (2004). *Multilevel Modeling.* Thousand Oaks: SAGE.

Mavletova, A. (2013). Data Quality in PC and Mobile Web Surveys. *Social Science Computer Review, 31*(6), 725-743. doi:10.1177/0894439313485201

Mayerl, J., Sellke, P., & Urban, D. (2005). *Analyzing Cognitive Processes in CATI-Surveys with Response Latencies: An Empirical Evaluation of the Consequences of Using Different Baseline Speed Measures.* Schriftenreihe des Instituts für Sozialwissenschaften der Universität Stuttgart. Universität Stuttgart. Stuttgart.

Olson, K., & Peytchev, A. (2007). Effect of Interviewer Experience on Interview Pace and Interviewer Attitudes. *Public Opinion Quarterly, 71*(2), 273-286. doi:10.1093/poq/nfm007

Peytchev, A., & Hill, C. A. (2010). Experiments in Mobile Web Survey Design: Similarities to Other Modes and Unique Considerations. *Social Science Computer Review, 28*(3), 319-335. doi:10.1177/0894439309353037

Rabe-Hesketh, S., & Skrondal, A. (2008). *Multilevel and Longitudinal Modeling Using Stata*. College Station: Stata Press.

Rada, V. D. d., & Dominguez-Alvarez, J. A. (2013). Response Quality of Self-Administered Questionnaires: A Comparison Between Paper and Web Questionnaires. *Social Science Computer Review, 32*(2), 256-269. doi:10.1177/0894439313508516

Rattinger, H., Roßteutscher, S., Schmitt-Beck, R., Schoen, H., Weßels, B., & Wolf, C. (2009-2015). *Long-term Online Tracking, T1-T29 (ZA5334-ZA5351 and ZA5719-ZA5729).* Cologne: GESIS Data Archive.

Reja, U., Manfreda, K. L., Hlebec, V., & Vehovar, V. (2003). Open-Ended vs. Close-Ended Questions in Web Questionnaires. *Developments in Applied Statistics, 19*, 159-177.

Revilla, M., & Ochoa, C. (2015a). Open Narrative Questions in PC and Smartphones: Is the Device Playing a Role? *Quality & Quantity*, 1-19.

Revilla, M., & Ochoa, C. (2015b). What are the Links in a Web Survey Among Response Time, Quality, and Auto-Evaluation of the Efforts Done? *Social Science Computer Review, 33*(1), 97-114. doi:10.1177/0894439314531214

Roßmann, J. (2017). *Satisficing in Befragungen. Theorie, Messung und Erklärung [Satisficing in Surveys. Theory, Measurement, and Explanation].* Wiesbaden: Springer VS.

Roßmann, J., & Gummer, T. (2016). *PARSEUAS: Stata Module to Extract Detailed Information From User Agent Strings.* Chestnut Hill, MA: Boston College.

Roßmann, J., Gummer, T., & Silber, H. (2018). Mitigating Satisficing in Cognitively Demanding Grid Questions: Evidence from Two Web-Based Experiments. *Journal of Survey Statistics and Methodology, 6*(3), 376-400. doi:10.1093/jssam/smx020

Rothman, K. J. (1990). No Adjustments are Needed for Multiple Comparisons. *Epidemiology*, 43-46.

Schuman, H., & Presser, S. (1996). *Questions and Answers in Attitude Surveys*. New York: Academic Press.

Schumann, S., & Schoen, H. (2009). Muster an Beständigkeit? Zur Stabilität politischer und persönlicher Prädispositionen. In S. Huber, H. Rattinger, & C. Wagner (Eds.), *Vom Interview zur Analyse* (Vol. 1, pp. 13-34).

Silber, H., Lischewski, J., & Leibold, J. (2013). Comparing Different Types of Web Surveys: Examining Drop-Outs, Non-Response and Social Desirability. *Metodološki zvezki, 10*(2), 121-143.

Smyth, J. D., Dillman, D. A., Christian, L. M., & McBride, M. (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*(2), 325-337. doi:10.1093/poq/nfp029

Snijders, T., & Bosker, R. (1999). *Multilevel Analysis. An Introduction to Basic and Advanced Multilevel Modeling*. London: SAGE.

Stapleton, C. E. (2013). The Smartphone Way to Collect Survey Data. *Survey Practice, 6*(2), 1-7.

Toepoel, V., & Lugtig, P. (2014). What Happens if You Offer a Mobile Option to Your Web Panel? Evidence From a Probability-Based Panel of Internet Users. *Social Science Computer Review, 32*(4), 544-560. doi:10.1177/0894439313510482

Tourangeau, R., Couper, M. P., & Conrad, F. (2004). Spacing, Position, and Order. Interpretive Heuristics for Visual Features of Survey Questions. *Public Opinion Quarterly, 68*(3), 368-393. doi:10.1093/poq/nfh035

Tourangeau, R., & Rasinski, K. A. (1988). Cognitive Processes Underlying Context Effects in Attitude Measurement. *Psychological Bulletin, 103*(3), 299.

Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). *The Psychology of Survey Response*. Cambridge, New York, Melbourne, Madrid, Cape Town, Singapore, Sao Paulo: Cambridge University Press.

Turner, G., Sturgis, P., & Martin, D. (2015). Can Response Latencies Be Used to Detect Survey Satisficing on Cognitively Demanding Questions? *Journal of Survey Statistics and Methodology, 3*(1), 89-108. doi:10.1093/jssam/smu022

Yan, T., & Tourangeau, R. (2008). Fast Times and Easy Questions: The Effects of Age, Experience and Question Complexity on Web Survey Response Times. *Applied Cognitive Psychology, 22*(1), 51-68.

APPENDIX A

Table A.1 Survey participation statistics

		N			Completion Rate
Survey	Field Period	accepted invitation	screened out & rejected	breakoff	in %
1	Apr 30 - May 05, ’09	4557	803	442	88.2
2	Mai 27 - Jun 05, ’09	2566	945	409	74.8
3	Jul 03 - Jul 13, ’09	1820	272	415	73.2
4	Jul 31 - Aug 11, ’09	1927	176	607	65.3
5	Aug 24 - Sep 01, ’09	1879	228	512	69.0
6	Sep 18 - Sep 27, ’09	1634	268	213	84.4
7	Sep 29 - Oct 08, ’09	2163	623	393	74.5
8	Dec 10 - Dec 20, ’09	1803	275	397	74.0
9	Apr 15 - Apr 23, ’10	1563	222	205	84.7
10	Jun 24 - Jul 05, ’10	1671	290	243	82.4
11	Sep 16 - Sep 26, ’10	1858	586	124	90.3
12	Dec 09 - Dec 19, ’10	1636	357	135	89.4
13	Mar 09 - Mar 19, ’11	1604	246	221	83.7
14	May 23 - Jun 03, ’11	1618	185	283	80.3
15	Aug 24 - Sep 03, ’11	1643	316	169	87.3
16	Dec 08 - Dec 18, ’11	1640	303	223	83.3
17	May 02 - May 15, ’12	1709	427	266	79.3
18	Sep 17 - Oct 01, ’12	1517	254	188	85.1
19	Jan 04 - Jan 19, ’13	1532	326	172	85.7
20	May 24 - Jun 08, ’13	1626	350	228	82.1
21	Sep 09 - Sep 21, ’13	1373	184	177	85.1
22	Nov 29 - Dec 13, ’13	1648	384	215	83.0
23	Feb 21 - Mar 07, ’14	1493	265	205	83.3
24	May 09 - May 23, ’14	1446	199	203	83.7
25	Aug 29 - Sep 13, ’14	1404	231	162	86.2
26	Nov 21 - Dec 05, ’14	1446	174	253	80.1
27	Feb 27 - Mar 13, ’15	1375	165	181	85.0
28	Jun 05 - Jun 19, ’15	1569	388	162	86.3
29	Sep 11 - Sep 25, ’15	1460	282	151	87.2

APPENDIX B

Operationalization of Respondent Level Variables

This appendix describes the operationalization of the respondent level variables that rely on the questions asked in 29 surveys. These variables include education, age, interest in politics, strength of a respondent's identification with a political party, (intended) turnout to vote at a federal election, sex, and region of residence.

Education

We categorized respondents' formal education as *low, intermediate*, and *high*. Since the response options to the open-ended question regarding respondents' formal level of education slightly changed throughout the 29 surveys, we relied on a standardized scheme of coding. The qualification that enabled students to enter a university was coded as *high* education while completing secondary/high school was considered to be an *intermediate* education. Anything less than completing secondary/high school was categorized as *low* education. For analytical purposes, we treated the variable as a categorical variable (0 = low / 1 = intermediate / 2 = high).

Age

According to their age, we coded respondents in five categories: 18–29 years, 30–39 years, 40–49 years, 50–59 years, and 60 years or older. Age was measured differently throughout the surveys. In surveys 1–7, respondents had to select one of the following categories in a close-ended question: 18–29 years, 30–39 years, 40–49 years, 50–59 years, and 60 years and above. Since survey 8, respondents have been asked about their date of birth in an open-ended question. For analytical purposes, we treated the variable as a categorical variable (0 = 18–29 / 1 = 30–39 / 2 = 40–49 / 3 = 50–59 / 4 = 60+).

Interest in the survey topic

We measured respondents' interest in the survey topic by a question on their political interest. This question used a 5-point scale that was labeled *very strong, fairly strong, moderately, fairly weak*, and *very weak*. We recoded the answers *very strong* and *fairly strong* as high political interest; the answer *moderately* as intermediate political interest; and the responses *fairly weak* and *very weak* as low political interest. Accordingly, respondents' interest in politics was coded as *low, intermediate*, and *high*. Again, for analytical purposes, we treated this variable as a categorical variable (0 = low / 1 = intermediate / 2 = high interest in politics).

Strength of party identification

We asked a question regarding the strength of a respondent's political party identification once they had stated they identified with a political party in a previous question. They had to answer the question on a 5-point scale that was labeled *very strong, fairly strong, moderately, fairly weak*, and *very weak*. If respondents did not identify with a political party, we coded their strength of party identification as *none*. If respondents identified with a party and reported the strength to be *fairly weak, very weak*, or *moderately*, we considered this as *moderate* strength. We coded respondents with a party identification of *fairly strong* or stronger as *strong*. For analytical purposes, we considered this variable as a categorical variable (0 = none / 1 = moderate / 2 = strong).

Intention to vote

To investigate respondents' motivation to participate in an election, we differentiated between the respondents who intended to turn out to vote at a federal election and those who did not. All surveys, except survey 7, featured a question on whether respondents would take part in the next German federal election. The five response options were *certain to vote, likely to vote, might vote, likely not to vote,* and *certain not to vote*. We coded respondents that reported to be *certain* or *likely* to vote as *yes*, while we considered the other respondents to have *no* intention to turn out. In survey 7, which was fielded in the aftermath of the German federal election 2009, a question on the actual turnout (*yes* or *no*) was asked. We used this question to code respondents of survey 7 either as *yes* or *no* with respect to their intention to vote at an election. For analytical purposes, we considered this variable as a binary variable (0 = no / 1 = yes).

Sex

We asked respondents about their sex with the response options *male* and *female*. For analytical purposes, we created a binary variable (0 = male / 1 = female).

Region of residency

We asked the respondents in which federal state of Germany they currently were residing. We coded the federal states Schleswig-Holstein, Hamburg, Lower Saxony, Bremen, North Rhine-Westphalia, Hesse, Rhineland-Palatinate, Baden-Wuerttemberg, Bavaria, and Saarland as *West Germany*; and we coded Berlin, Brandenburg, Mecklenburg-Vorpommern, Saxony, Saxony-Anhalt, and Thuringia as *East Germany*. For analytical purposes, we created a binary variable (0 = East Germany / 1 = West Germany).

methods, data, analyses | Vol. 14(1), 2020, pp. 35-60 DOI: 10.12758/mda.2020.01

Interviewer Training Guidelines of Multinational Survey Programs: A Total Survey Error Perspective

Daniela Ackermann-Piek[1], *Henning Silber*[1], *Jessica Daikeler*[1], *Silke Martin*[1] *& Brad Edwards*[2]

[1] *GESIS – Leibniz Institute for the Social Sciences, Germany*

[2] *Westat, Rockville, Maryland, USA*

Abstract

Typically, interviewer training is implemented in order to minimize interviewer effects and ensure that interviewers are well prepared to administer the survey. Leading professional associations in the survey research landscape recommend the standardized implementation of interviewer training. Some large-scale multinational survey programs have produced their own training guidelines to ensure a comparable level of quality in the implementation of training across participating countries. However, the length, content, and methodology of interviewer training guidelines are very heterogeneous. In this paper, we provide a comparative overview of general and study-specific interviewer training guidelines of three multinational survey programs (ESS, PIAAC, SHARE). Using total survey error (TSE) as a conceptual framework, we map the general and study-specific training guidelines of the three multinational survey programs to components of the TSE to determine how they target the reduction of interviewer effects. Our results reveal that unit nonresponse error is covered by all guidelines; measurement error is covered by most guidelines; and coverage error, sampling error, and processing error are addressed either not at all or sparsely. We conclude, for example, that these guidelines could be an excellent starting point for new – small as well as large-scale – surveys to design their interviewer training, and that interviewer training guidelines should be made publicly available in order to provide a high level of transparency, thus enabling survey programs to learn from each other.

Keywords: interviewer training guidelines, interviewer effects, multinational survey programs, total survey error

Concerns about interviewer effects in interviewer-mediated surveys have accompanied generations of survey researchers. Most of the literature on interviewer effects focuses on the description and explanation of these effects after data collection (West & Blom, 2017). However, in order to ensure that interviewer-administered surveys produce high-quality data, it is essential that measures be taken to prevent, or minimize, interviewer effects. One such measure is the implementation of standardized interviewer training. In addition, interviewer training is typically used to ensure that interviewers are well prepared to adequately perform all the tasks they have during the survey implementation.

To date, findings on the effects of interviewer training on data quality are quite heterogeneous. Although most studies have shown large positive effects of interviewer training (e.g., Benson & Powell, 2015; Billiet & Loosveldt, 1988; Fowler Jr & Mangione, 1986; Fowler Jr., 1991; Mayer & O'Brien, 2001), some have found only small positive effects (e.g., Cantor, Allen, Schneider, Hagerty-Heller, & Yuan, 2004; McConaghy & Carey, 2005), and others have failed to identify any significant positive effects (e.g., Schnell & Trappmann, 2006). One reason for the heterogeneity of the findings on the effects of interviewer training may be that the training programs themselves are very heterogeneous (for a short overview, see Daikeler & Bosnjak, forthcoming; Lessler, Eyerman, & Wang, 2008). Thus, it is not surprising that general interviewer training have been proposed from multiple sources, all recommending a careful planning and standardized implementation of interviewer training when conducting large-scale interviewer-administered surveys (American Association for Public Opinion Research [AAPOR], n.d.; Alcser, Clemens, Holland, Guyer, & Hu, 2016; Daikeler, Silber, Bosnjak, Zabal, & Martin, 2017; Fowler Jr. & Mangione, 1990; International Organization for Standardization [ISO], 2012; Lessler et al., 2008). Interviewer training is part of the training concepts of large-scale interviewer-administered survey programs, such as the European Social Survey (ESS), the Survey of Health, Ageing and Retirement in Europe (SHARE), and the Programme for the International Assessment of Adult Competencies (PIAAC), and of the vast majority of large-scale national interviewer-administered surveys, such as the German Socio-Economic Panel (SOEP) and the U.S. National Health

Acknowledgments

The authors thank the participants of the workshop „Dashboard & Field Monitoring" held in Mannheim, 2018 for their thoughts which encouraged this research idea. We are grateful to Sabine Friedel and Verena Halbherr for helpful details about interviewer training in the ESS and SHARE and detailed cross-checks of the manuscript.

Direct correspondence to

Daniela Ackermann-Piek, GESIS – Leibniz Institute for the Social Sciences,
Quadrat B2 1, 68159 Mannheim, Germany
E-mail: daniela.ackermann-piek@gesis.org

Interview Survey (NHIS). More specifically, in line with the general guidelines issued by leading professional associations in the area of survey research, the interviewer training concepts of national and multinational survey programs include recommendations for study-specific training for inexperienced and experienced interviewers, as well as brief sections on general interviewer training for inexperienced interviewers.

However, what is lacking in the literature is a structured comparison of the content of the various interviewer training guidelines of survey programs or training concepts of large surveys – using a theoretical framework. The present article aims to fill this gap by providing a comparative overview of the extent to which the content of training guidelines of the ESS, PIAAC, and SHARE (Börsch-Supan & Jürgens, 2005; ESS, 2016a, 2016b, Malter & Börsch-Supan, 2017; Organziation for Economic Co-operation and Development [OECD], 2011, 2013) are integrated into the conceptual *total survey error* framework (TSE; Biemer, 2010; Groves et al., 2009; Pennell, Hibben, Lyberg, Mohler, & Worku, 2017; T. W. Smith, 2019). Additionally, we investigate the extent to which the individual components of the TSE are addressed in these guidelines, and we make suggestions for improvements. Specifically, we focus on the topics of interviewer training specified in training guidelines on an international level rather than on interviewer training in specific countries with detailed training content.

The TSE Framework and the Literature on Interviewer Training

The TSE is a theoretical concept that describes statistical error properties of survey estimates, systematically structured along different error sources (Biemer, 2010; Groves & Lyberg, 2010). Error sources are assigned to each step in the survey life cycle, typically along two dimensions each, either sampling error and nonsampling error (Biemer, 2010) or measurement and representativeness (Groves et al., 2009). As our aim is to compare interviewer training of multinational survey programs along the TSE, we follow the approach of Pennell et al. (2017), who adopted the TSE typology for multinational, multiregional, and multicultural surveys (3MC). Pennell et al. (2017) provide a TSE model which combines the complexity in designing and implementing 3MC surveys with the overall aim to minimize comparison error (T. W. Smith, 2011). Following the approach of Groves et al. (2009), the authors structure their model along the two dimensions measurement and representativeness.

Following Groves et al. (2009) and Pennell et al. (2017), the representation dimension of the TSE includes *coverage error*, *sampling error*, *nonresponse error*, and *adjustment error* and the measurement dimension includes *validity*,

measurement error, and *processing error*. *Coverage error* refers to problems of a not perfectly covered target population in the sampling frame. *Sampling error* occurs because only a sample is observed instead of the entire target population. With regard to sampling error, either bias (members of the sampling frame are systematically excluded from selection) or variance (different sets of sample frame elements are selected and each set can have different values in the survey statistic) can occur. *Nonresponse error* occurs when selected sample members that respond the survey request systematically differ from those who do not respond the survey request. After data collection, post-survey adjustments are typically used to correct for representation errors occurred earlier in the process. However, when post-survey adjustments fail to capture each case of misrepresentation in the sample, *adjustment error* occurs. The error components of the measurement dimension of the TSE are associated with errors in the survey instruments and the question-response process. The first error component of the measurement dimension of the TSE, *validity*, reflects an error that describes that the theoretical construct is not optimally reflected in the measure. The next error component, *measurement error,* occurs when the response given by a respondent differs from the true response. Finally, *processing error* reflects the incorrect transfer of responses to data storage during capturing, coding, or editing of data. All errors described in the TSE may result in biased survey estimates of substantive survey variables. Thus, the aim of survey operations is to minimize the errors under the given time and cost constraints to maximize the survey quality (Schouten, Peytchev, & Wagner, 2017).

The TSE framework is regularly used to describe and structure interviewer errors in interviewer-administered surveys (for an overview, see West & Blom, 2017). As interviewers have many tasks when administering a survey, such as contacting sample units, gaining their cooperation, asking questions, and recording answers (e.g., Groves et al., 1992; Loosveldt, 2008; Schaeffer, Dykema, & Maynard, 2010), they can – intentionally or unintentionally – affect a large number of steps in the survey process. In other words, they can be the sources of multiple survey errors. Because interviewer training is organized along interviewer tasks, the TSE framework can be used to structure training content when reviewing general and study-specific interviewer training concepts of multinational survey programs.

A review of the literature on the effects of interviewer training on survey data quality from a TSE perspective reveals that most studies focus either on nonresponse error or measurement error. Studies on the effects of interviewer training on nonresponse error, for example, show a positive effect of refusal avoidance training on reducing nonresponse (Cantor et al., 2004; Daikeler & Bosnjak, forthcoming; Durand, Gagnon, Doucet, & Lacourse, 2006; Hubal & Day, 2006; Mayer & O'Brien, 2001; O'Brien, Mayer, Groves, & O'Neill, 2002). However, not all studies have found overall positive effects of interviewer training on nonresponse (Schnell & Trappmann, 2006). Studies with a focus on the effects of interviewer training

on several quality indicators related to measurement error, for example, have identified a positive effect of interviewer training on the application of standardized interviewing techniques (Billiet & Loosveldt, 1988; Dahlhamer, Cynamon, Gentleman, Piani, & Weiler, 2010; Fowler Jr., 1991), the reduction of item nonresponse (Billiet & Loosveldt, 1988; Daikeler & Bosnjak, forthcoming), and the application of appropriate probing techniques (Billiet & Loosveldt, 1988; Daikeler & Bosnjak, forthcoming). However, Groves (2005) noted that the literature left open the question whether interviewer training effectively reduces measurement error. And finally, the effects of interviewer training on coverage error, sampling error, and processing error have been addressed only occasionally in the literature (Eckman & Kreuter, 2011; Guest, 1954).

Methodology and Resources

The present study compares interviewer training concepts of three large-scale multinational survey programs in the social sciences, namely, the ESS (ESS, 2016a, 2016b, 2018), PIAAC (OECD, 2011, 2013), and SHARE (Börsch-Supan & Jürges, 2005; Malter & Börsch-Supan, 2017). In most cases, only multinational survey programs have the funds to develop detailed interviewer training guidelines and implement interviewer training accordingly. These programs need predefined detailed specifications for participating countries, because such programs are imperative to ensure a harmonized data collection process across countries, which is a prerequisite for obtaining high-quality data (Pennell, Harkness, Levenstein, & Quaglia, 2010; Survey Research Center at the University of Michigan, 2016). Ensuring harmonization across countries also applies to the training of interviewers.

Overall, the ESS, PIAAC and SHARE fulfilled the following selection criteria: First, all three are administered by interviewers face-to-face. Second, they have participants from many European countries. Third, in all three cases detailed documents were publicly available that contained information on the survey programs' interviewer training guidelines. Very often this information is confidential and not accessible.

The ESS is a cross-sectional survey of attitudes, beliefs, and behaviors that is conducted every two years. SHARE is a longitudinal survey on health, ageing and retirement. SHARE is also conducted every two years. For these two survey programs, we selected the specifications and characteristics from the last round for which interviewer training guidelines are available (ESS Round 8, 2016; SHARE Wave 6, 2015). PIAAC is a multi-cycle program for the assessment of basic adult

competencies; a cross-sectional "Survey of Adult Skills" is carried out every 10 years. We used the training material from PIAAC Cycle 1, Round 1, 2011/2012[1].

Training concepts relate to training guidelines on an international level rather than to interviewer training in specific countries. Comparing the implementation of interviewer training in the various participating countries in detail would be another important research question. Also, we focus our research on general as well as survey specific interviewer training content provided by the three multinational survey programs.

Interviewer Tasks within ESS, PIAAC, and SHARE

When comparing interviewer training guidelines of different survey programs, it is important to take interviewers' tasks in the surveys and the resulting complexity of their roles into account. First of all, this refers to the target population, as these are the persons with whom interviewers interact. The target population of the ESS and PIAAC is quite similar and refer to the general population aged either 15 years or older (ESS) or between 16 and 65 years (PIAAC). The target population in SHARE also refers to the general population, however, only to persons who are 50 years or older at the time of sampling. In addition, in SHARE, spouses or partners of the sampled person are interviewed as well, if applicable.

Another interviewers' task for all three survey programs was the administration of the core questionnaire face-to-face using computer-assisted personal interviewing (CAPI). In addition, PIAAC and SHARE interviewers had to perform additional tasks. In PIAAC, interviewers had to administer a cognitive assessment where respondents worked independently on a number of tasks on the interviewer's laptop or in a paper booklet under the supervision of the interviewer. For this purpose, interviewers switched from their traditional role of asking questions and took on a passive, monitoring role, adapting their behavior accordingly. If the respondent opted for the paper-based cognitive assessment, the interviewer additionally had to score some items for routing purposes.[2] In SHARE, interviewers administered a self-completion paper questionnaire to the respondents in some countries. As the target population in SHARE consisted of elderly persons, interviewers had to be able to interact with this special population. A special and new task for some SHARE interviewers was to collect biomarkers from respondents and conduct physical tests (e.g., measuring blood pressure). The average interview duration was

1 For a more detailed overview of the specifications and characteristics of the ESS, PIAAC, and SHARE across all participating countries, see the Appendix.

2 Scoring means that the interviewer has to determine a value (correct or incorrect) for each response to a number of selected items based on scoring guidelines (Zabal et al., 2014, p. 104).

60 minutes for the ESS (Round 8), whereas SHARE and PIAAC had longer average durations (80 minutes and 90 minutes, respectively).

Interviewer Training Concepts of ESS, PIAAC, and SHARE

The ESS specifications for the countries distinguish between two types of interviewer preparation: training and briefing. ESS interviewers are expected to have previous face-to-face interviewing experience and to be trained in effective doorstep interaction, standardized interviewing techniques, and general interviewer behavior before administering the survey instrument. In each round of the ESS, experienced interviewers receive a briefing, whereas inexperienced interviewers should additionally undergo general interviewer training prior to the briefing.

With respect to interviewer training in PIAAC, several features can be highlighted: (1) the extensive interviewer training package (including, e.g., fully scripted training sessions); (2) the train-the-trainer session prior to national interviewer training in which the training staff is introduced to the scripts and interview materials; (3) the close monitoring by the international consortium of the implementation of the country-based interviewer training. As PIAAC Round 1, Cycle 1 also included a field trial in which all aspects of the survey – including interviewer-related topics – were tested, the interviewer training sessions for the main study were shortened depending on the performance of the interviewers in the field trial.

For SHARE, the survey programs' multiplier approach to interviewer training can be highlighted: a centralized train-the-trainer program is conducted to facilitate decentralized interviewer training in the participating countries. Moreover, all interviewers are expected to have extensive general face-to-face interviewing experience and to have received in-person general interviewer training prior to undergoing study-specific training.

Standards for the Implementation of Interviewer Training within ESS, PIAAC, and SHARE

The extent to which the implementation of interviewer training is specified differs considerably across the three multinational survey programs (for an overview, see Table 1). As a first impression, when counting the number of pages in the overall survey specifications which are provided to the participating countries for the respective survey[3], it becomes obvious that the specifications for the ESS (65 pages)

3 The survey specifications are provided to the country contact and are intended to be used as orientation for the implementations of the survey. Typically, these specifications are not handed out to the interviewers.

Table 1 ESS, PIAAC, and SHARE Standards for the Implementation of Interviewer Training and Desired Interviewer Characteristics

	ESS 2016 (Round 8)[1]	PIAAC (Cycle 1, Round 1)[2]	SHARE (Wave 6)[3]
General concept	Interviewer experience important: GIT for inexperienced interviewers and reduced training for experienced interviewers	Extensive material provided for standardized training sessions for all interviewers Training for pretest and main study (the latter can be reduced depending on experience and performance in pretest)	Multiplier approach: Centralized TTT program to facilitate decentralized national training Interviewers expected to be experienced
Length of survey specifications (number of pages)	65	199	542
Material for trainers	NC manual Series of pre-structured slides with movie clips and related material (incl. guidelines for training, scripted practice interview)	Technical Standards and Guidelines manual for NPM teams (plus comprehensive interviewer training material on planning and implementation, incl. PowerPoint slides, training scripts)	Interviewer project manual Facilitator guide (incl. PowerPoint slides, training scripts) CD-based training on gaining respondent cooperation (incl. training videos) Computer-based tutorial (using SHARE CMS) Training evaluation protocols Sample management monitoring Specific guidelines (to be followed by interviewers to ensure cross-national comparability)

	ESS 2016 (Round 8)[1]	PIAAC (Cycle 1, Round 1)[2]	SHARE (Wave 6)[3]
Material for interviewers	Interviewer manual	Home-study material, incl. written exercises Interviewer manual	Interviewer manual (incl. instructions for blood collection with a short video)
Training facilities and equipment	U-shaped chair setup, technical equipment	Adequate space (separate rooms, each with a lead trainer, technical equipment)	---
Training staff	---	Competent and experienced lead trainers Assistant trainers Technical support (attended TTT sessions)	International trainer trains the national trainers
Supervisory staff	---	Regional supervisors and field managers who attended supervisor training session	---
Training delivered to:	Interviewers	Interviewers Trainers Supervisors Field managers	National trainer (from the survey agency) Country team leader Operator Interviewers
Interviewer experience	Experienced, trained in general interviewer tasks F2F Reduced training (“briefing”) for all interviewers, general training for new interviewers	New interviewers need GIT/training in CAI prior to/during study-specific training	New interviewers need GIT Extensive general face-to-face interviewing experience

	ESS 2016 (Round 8)[1]	PIAAC (Cycle 1, Round 1)[2]	SHARE (Wave 6)[3]
Mode of training	In-person	In-person	In-person
Trainee group size	---	15–20	---
Training length	6 hrs	Pretest: 36 hrs all interviewers Main study: 15 hrs (pretest interviewers with good reviews); 30 hrs (pretest interviewers with less than favorable reviews or interviewers with experience on other surveys), 36 hrs inexperienced interviewers	TTT pilot: 2 days TTT pretest 1.5 days TTT main: 1 days national training: 2 days
Training scheduling	One month before fieldwork starts	No earlier than two weeks before, but preferably the week immediately before the scheduled start of data collection	TTT pilot: --- TTT pretest: --- TTT main: --- National training: Dec. 2014 – Feb. 2015
Training methods	Presentation Mock interviews Video clips (scenarios) Role play Interactive discussion	Lectures Scripted mock interviews Roundtable exercises Role play Practice interviews Training in multiple languages (if applicable)	Mock interviews Question cards Question-answer session Practice for blood collection
Training evaluation	---	Interviewers' performance (during training) Supervisors' ability to perform responsibilities	Local survey agency trainers evaluate TTT in writing, preparation for local training sessions, materials, and possible improvements

	ESS 2016 (Round 8)[1]	PIAAC (Cycle 1, Round 1)[2]	SHARE (Wave 6)[3]
Quality control	NC must oversee interviewer selection and training and make sure that in-person training takes place	Documentation of plan, scheduling, material, implementation, and evaluation of national training Training report Retraining and attrition training during data collection (if applicable)	Trainer certification Interviewer certification

Note. Survey programs in alphabetical order; information retrieved from general project guidelines or project specifications (see references below); CAI = computer-assisted interviewing; CMS = case management system; ESS = European Social Survey; F2F = face-to-face; GIT = general interviewer training; NC = National Coordinator; NPM = National Project Management; PIAAC = Programme for the International Assessment of Adult Competencies; SHARE = Survey of Health, Ageing and Retirement in Europe; TTT = train-the-trainers session; --- = no information included in the specifications.

[1] ESS (2016b); ESS (2016a); ESS (2018).
[2] OECD (2011); OECD (2013, Chapter 10.4).
[3] Börsch-Supan and Jürges (2005); Malter, Schuller, and Börsch-Supan (2016).

are not as extensive as those for PIAAC (199 pages) and SHARE (542 pages). This is not surprising when one considers the additional, non-standard, tasks that interviewers in PIAAC and SHARE must perform. In addition, experienced ESS interviewers typically only receive a reduced version of the interviewer training (referred to as "interviewer briefing").

The interviewer training guidelines of all three survey programs specify that the training should be conducted in-person. All guidelines require measures for controlling the quality of the training (e.g., review of interviewer selection and training report by country). In contrast, other topics are not covered by the survey specifications of all three survey programs. For example, the ESS and PIAAC specify that training should be scheduled to take place shortly before the start of data collection, whereas this is not addressed in the SHARE survey specifications. Moreover, the ESS and SHARE specifications related to interviewer training do not address training of supervisory staff, training-group size, or the structure of training groups, whereas PIAAC defines these aspects clearly. In addition, the SHARE specifications do not address training facilities and equipment, and the ESS specifications do not include information on the evaluation of the interviewer training. Other examples are that in the ESS, for example, the national coordinators, who are responsible for organizing the national interviewer training, are given research-based information on interviewer effects to demonstrate the positive effects of interviewer training. The SHARE specifications contain information about the national interviewer training and the train-the-trainer sessions and the PIAAC specifications emphasize the importance of quality controls and provide very detailed guidelines on the implementation of interviewer training.

Mapping of Interviewer Training Content to the TSE

In this section, we map the specifications for general interviewer training and the program-specific training content of the three multinational survey programs along the TSE framework.

General Interviewer Training for Inexperienced Interviewers

Table 2 provides an overview of training content of the interviewer training guidelines of the three multinational survey programs from a TSE perspective for general interviewer training, intended for inexperienced interviewers. Of the three survey programs, only PIAAC provides comprehensive guidelines on general interviewer

training for inexperienced interviewers, which go beyond the coverage of nonresponse and measurement error. In contrast, the ESS and SHARE interviewer training guidelines include only some examples of topics covering interviewer training for inexperienced interviewers.

Table 2 ESS, PIAAC, and SHARE General Interviewer Training Content from a TSE Perspective for Inexperienced Interviewers

	ESS[1]	PIAAC[2]	SHARE[3]
Representation			
Coverage error	---	---	---
Sampling error	---	---	---
Nonresponse error	Doorstep interaction	Gaining cooperation (incl. detailed contact and refusal conversion strategies)	Collecting process data information (incl. contact attempt and result of contact attempt)
Measurement			
Measurement error	Standardized interviewing (incl. detailed rules)	Asking questions (incl. exercises) Probing techniques	Standardized question-asking Probing
Processing error	---	Recording answers (incl. exercises)	---
Content of relevance for multiple TSE components			
	---	Introduction to survey research (incl. types of survey questions, interviewing terminology) Standards and ethics in survey research (incl. informed consent, data confidentiality, data security, exercises) Remuneration and administrative aspects Basics of computer-assisted interviewing (CAI)	---

Note. Survey programs in alphabetical order; ESS = European Social Survey; PIAAC = Programme for the International Assessment of Adult Competencies; SHARE = Survey of Health, Ageing and Retirement in Europe; --- = no information included in the guidelines.

[1] ESS (2016c); Beullens, Loosveldt, Denies, and Vandenplas (2016).

[2] OECD (2013).

[3] Börsch-Supan and Jürges (2005).

The PIAAC specifications for general interviewer training, that affect all components of the TSE, comprise an introduction, some standards and ethics in survey research, administrative aspects, and instructions on the basics of computer-assisted interviewing (CAI). With regard to nonresponse, these specifications include strategies for gaining cooperation, and with regard to measurement error, they relate to question-answering and probing techniques. In addition, processing error is covered by techniques for recording answers. Nevertheless, looking at the proposed length of the training sessions for each topic, it becomes apparent that the focus is clearly on measurement error and on related practice sessions (OECD, 2013, Chapter 10.4).

In contrast, the ESS specifications for general interviewer training cover only nonresponse error (doorstep interaction training) and measurement error (training of standardized interviewing). Similarly, the SHARE specifications for general interviewer training are quite brief and cover only nonresponse error (training in process data collection) and measurement error (standardized interviewing and probing techniques). Neither the ESS nor the SHARE specifications include information about the length of the general interviewer training.

Program-Specific Interviewer Training for Inexperienced and Experienced Interviewers

In Table 3, we map the program-specific training content of the three survey programs to the components of the TSE. Training content that is relevant for all error sources is presented at the bottom of the table. Program-specific training is intended for interviewers who have general interviewing experience or who have attended general interviewer training but are not familiar with the program-specific interview tasks.

Table 3 Program-Specific Interviewer Training Content of the ESS, PIAAC, and SHARE from a TSE Perspective for all Interviewers

	ESS[1]	PIAAC[2]	SHARE[3]
Representation			
Coverage error	Selecting respondents	Screener administration	
Sampling error	---	---	---
Nonresponse error		Screener administration	
		Locating strategy	Locating strategy
	Contact strategy (detailed)	Contact strategy	Contact strategy (detailed)
	Gaining cooperation (incl. realistic examples)	Gaining cooperation	Gaining cooperation (detailed, focus: representativeness)
	Refusal conversion	Refusal avoidance and conversion	Refusal conversion
Measurement			
Measurement error		Screener administration	
	Instrument overview (incl. practice, focus: specific questions)	Instrument administration (incl. practice)	Instrument overview (incl. practice)
			Probing
	Standardized interviewing		
			Collection of a dried blood sample (detailed)
Processing error	---	---	Coding conventions (Mental Health section)
			Recording responses

	ESS[1]	PIAAC[2]	SHARE[3]
Content of relevance for multiple TSE components			
	Introduction (incl. goal, findings previous rounds, data quality issues, data usage)	Introduction	Introduction (incl. goal, questionnaire overview)
	Logistics (incl. target response rate, *supplementary material*, *remuneration*, fieldwork procedures, etc.)		Logistics
		Review of advanced materials	
		Case management system (incl. data transmission)	Laptop overview (incl. installation check)
	Administrative tasks (contact forms)	Administrative tasks (incl. disposition codes, case folder, record of contact)	Case management system (incl. practice)
	Information for respondents prior interview (e.g., data protection; general information ESS; data confidentiality; data storage, etc.)		
	Experienced interviewers: changes since last round, comparison with other surveys	*Demo* and practice interviews (role play and/or live respondent practice)	
		Question-and-answer session (after practice)	Question-and-answer session
		Additional practice sessions	
	Quality control	Quality control and monitoring	

Note. Survey programs in alphabetical order; ESS = European Social Survey; PIAAC = Programme for the International Assessment of Adult Competencies; SHARE = Survey of Health, Ageing and Retirement in Europe; --- = no information included in the specifications; optional training content in italics.

[1] ESS (2016c); Beullens et al. (2016).

[2] OECD (2013).

[3] Börsch-Supan and Jürges (2005); Malter and Börsch-Supan (2017).

A comparison of the training content of the ESS, PIAAC, and SHARE reveals that the survey programs focus on different components of the TSE framework. However, some training content is similar for all three survey programs, for example, an overall introduction to the survey, which is relevant for multiple components of the TSE. In addition, all three survey programs offer training content on contacting, gaining cooperation, and refusal avoidance strategies, with the goal of reducing nonresponse error. To address measurement error, the interviewer training specifications of all three survey programs include sessions providing an overview of the survey instruments.

The differences between the training guidelines of the three survey programs reveal that the ESS specifications include very precise information on survey logistics (e.g., target response rate, fieldwork procedures), administrative tasks, and information that must be provided to respondents before the interview starts (e.g., data confidentiality, data storage). These topics included in the ESS specifications are relevant for all TSE components as it might affect more than one error source. Similarly, the specifications for interviewer training in PIAAC are quite detailed with respect to administrative tasks, and they additionally include a large section on practical sessions (e.g., question-and-answer sessions, demo interviews), which are also relevant for multiple error sources. In comparison, SHARE does not include detailed specifications for administrative tasks. However, in the training session on mental health, there is a large sub-section on coding conventions. Training in coding conventions aims to reduce processing error; this is not covered by the specifications of the other two survey programs. Yet, the SHARE specifications for interviewer training do not include any information on quality control and monitoring, which is covered by the ESS and PIAAC training content specifications.

Discussion

In the present paper, we aimed to review program-specific interviewer training guidelines of three multinational large-scale survey programs (ESS, PIAAC, and SHARE) using the TSE framework. Our results show that there is a clear focus on measurement error, nonresponse error, and introductory and administrative topics in the training materials. Other error sources are either covered by more general parts of the interviewer training guidelines (e.g., logistics, technical issues), which address multiple components of the TSE framework, or are rarely (e.g., processing error is covered only by the SHARE interviewer training guidelines) or not covered at all. There are several possible explanations for this. First, it is reasonable that the focus of the training reflects the actual tasks of interviewers: gaining cooperation and the administration of the question-and-answer process are among an interviewer's main tasks in almost every survey program. The tasks assigned to

interviewers vary largely across different survey projects. Thus, the involvement of interviewers related to tasks affecting coverage, sampling, and processing is not part of every survey project, as in the three cases examined here. Second, the measurement and detection of coverage, sampling, and processing errors requires a higher control effort compared to the two other error sources. And third, most surveys have experienced a dramatic decrease in response rates in recent years (Beullens, Loosveldt, Vandenplas, & Stoop, 2018), which might make the skill of gaining cooperation more salient.

When looking at the interviewer training guidelines of the three survey programs in more detail, we identified several differences between the training content of the three multinational training programs. Training content aimed at reducing nonresponse error was identified in the interviewer training guidelines of all three multinational survey programs. However, training in locating sample units is not mentioned in the ESS guidelines. Training content aimed at reducing measurement error is covered in all three guidelines: the ESS guidelines ar the only guidelines including standardized interviewing, whereas probing techniques and the collection of biomarkers or the administration of a cognitive test are included only in the SHARE or PIAAC guidelines. These differences are due, in part, to the scope of the respective studies, which obviously differs across the three surveys we have compared in the present paper. Besides, training content relating to processing error is covered only in the SHARE training guidelines. Moreover, the PIAAC interviewer training guidelines include an extensive general interviewer training agenda for inexperienced interviewers, which is only sparsely addressed by the ESS and SHARE. As PIAAC was conducted for the first time in the participating countries and – as general interviewer training forms the basis for additional project-specific training – we suspect that the interviewer training guidelines aimed to ensure that all interviewers working for PIAAC were at a similar level of knowledge. However, all three multinational survey programs require that only interviewers who are trained in general interviewer tasks are employed for the survey. This is in line with Pennell et al. (2017), as interviewers with more interviewing experience are likely to minimize comparison error.

An important limitation of our study is that we focus on general guidelines of multinational survey programs only, but do not include a country-level comparison. Although specific training content or formal aspects (e.g., the number of interviewers, interviewer payment) are defined in the guidelines of the survey programs, compliance with and implementation of these guidelines can vary considerably in the participating countries. In addition, in many countries survey agencies are responsible for organizing and conducting interviewer training. Arrangements between the survey agency and the national coordinator of the survey can also determine the content of the training. Therefore, examining the technical reports to compare the actual implementation of interviewer training at the country level

would be a promising avenue for future research. In addition, we use the TSE as theoretical framework for comparing training content of the three multinational survey programs. However, the TSE itself also has limitations. While using the TSE framework does allow us to map the recommended content, it does not allow us to map, for example, the recommended didactic methods.

Moreover, different learning strategies (e.g., class instructions and practical sessions) are implemented mainly in the case of training content related to measurement error. Specifically, all three guidelines highlight the importance of practical training sessions. For the development of interviewer training concepts, the application of different learning modes and methods appears to be an important aspect which has only be sparsely taken into account so far (e.g., Daikeler & Bosnjak, forthcoming; Rutgers Online Degrees, n.d.; M. K. Smith, 2002): the field of andragogy addresses this topic and offers principles that are useful to consider when designing interviewer training programs (Tusting & Barton, 2003). For example, following Malcolm Knowles, adults prefer a self-directed approach and learning that is centered around common tasks (Meuler, 2010; Rutgers Online Degrees, n.d.; M. K. Smith, 2002). Moreover, interviewer training materials should take account of the fact that levels of educational attainment and experience vary greatly among adults.

Looking forward, in order to empirically investigate the different effects of general and study-specific interviewer training on the components of the TSE and, thus, on data quality, more experimental studies are needed. These studies should explore the effects of the various interviewer training contents on the different error sources as well as the interaction of different error sources. For example, experimental evidence is needed to ensure that the focus on gaining cooperation, which is typical of many interviewer training concepts, contributes to effectively reducing nonresponse. A theoretical foundation could be the organizing model for future research investigating explanations for interviewer effects on multiple error sources, which West and Blom (2017) proposed in their research synthesis on interviewer effects. Their proposed model includes interviewer training as a background characteristic that can be modified depending on the sources of interviewer effects identified. Future studies could structure research topics of interviewer training along this model in order to evaluate their impact on the respective survey errors.

Conclusion

The study showed that the interviewer training guidelines of all three multinational survey programs provided an extensive training content that addresses multiple error sources of the TSE framework. While the coverage of some error sources such as sampling and processing error could be improved when interviewers are

involved in these processes, the most important error sources nonresponse and measurement error are broadly covered by all three training guidelines. Altogether, those guidelines could serve, together with the survey guidelines formulated outside of the context of specific survey programs (e.g., Alcser, et al., 2016; Daikeler, Silber, Bosnjak, Zabal, & Martin, 2017; ISO, 2012; Lessler et al., 2008), as an excellent starting point for new – small as well as large-scale – surveys to design their interviewer training. Even the interviewer training of existing multinational survey programs could benefit from learning how other surveys plan and implement interviewer training to ensure interviewers are well prepared for all tasks they have to fulfil during the implementation of the respective survey. It might be helpful to define the focus of interviewer training through determining the time devoted to a specific topic dependent on the magnitude of the survey error related to that training topic (West & Blom, 2017). For example, a focus on training in contact and cooperation strategies is undoubtedly a good strategy in times of lower response rates or higher nonresponse bias. However, other components of the TSE should be likewise addressed in the respective interviewer training guidelines.

All this is only possible when interviewer training guidelines and materials are publicly available. Consequently, all survey programs would benefit from a high level of transparency (e.g., published interviewer training material). And since not all survey programs can afford a cost intensive high quality interviewer training, it would be imperative to have a standardized, pre-established training manual from which even smaller surveys can use relevant training modules.

References

American Association for Public Opinion Research (AAPOR, n.d.). best practices for survey research. Retrieved from https://www.aapor.org/Standards-Ethics/Best-Practices.aspx (03/13/2018).

Alcser, K., Clemens, J., Holland, L., Guyer, H., & Hu, M. (2016). Interviewer recruitment, selection, and training. Guidelines for Best Practice in Cross-Cultural Surveys. Survey Research Center, Institute for Social Research, University of Michigan.

Benson, M. S., & Powell, M. B. (2015). Evaluation of a comprehensive interactive training system for investigative interviewers of children. Psychology, Public Policy, and Law, 21(3), 309-322. https://doi.org/10.1037/law0000052

Beullens, K., Loosveldt, G., Denies, K., & Vandenplas, C. (2016). Quality matrix for the European Social Survey, round 7. Retrieved from http://www.europeansocialsurvey.org/docs/round7/methods/ESS7_quality_matrix.pdf (2018/05/02)

Beullens, K., Loosveldt, G., Vandenplas, C., & Stoop, I. (2018). Response rates in the European Social Survey: Increasing, decreasing, or a matter of fieldwork efforts? Survey Methods Insights from the Field. Retrieved from https://surveyinsights.org/?p=9673

Biemer, P. P. (2010). Total survey error: Design, implementation, and evaluation. Public Opinion Quarterly, 74(5), 817-884. https://doi.org/10.1093/poq/nfq058

Billiet, J., & Loosveldt, G. (1988). Improvement of the quality of responses to factual survey questions by interviewer training. Public Opinion Quarterly, 52(2), 190–211. Retrieved from http://www.jstor.org/stable/2749273

Börsch-Supan, A., & Jürges, H. (2005). The Survey of Health, Aging, and Retirement in Europe – Methodology. Mannheim.

Cantor, D., Allen, B., Schneider, S. J., Hagerty-Heller, T., & Yuan, A. (2004). Testing an automated refusal avoidance training methodology. Paper presented at the 60th Annual Conference of the American Association for Public Opinion Research, Phoenix, AZ.

Dahlhamer, J. M., Cynamon, M. L., Gentleman, J. F., Piani, A. L., & Weiler, M. J. (2010). Minimizing survey error through interviewer training: New procedures applied to the National Health Interview Survey (NHIS).

Daikeler, J., & Bosnjak, M. (forthcoming). How to conduct effective interviewer training: A meta-analysis. In K. Olson, J. D. Smyth, J. Dykema, A. Holbrook, F. Kreuter, & B. T. West (Eds.), Interviewer Effects from a Total Survey Error Perspective: CRC Press.

Daikeler, J., Silber, H., Bosnjak, M., Zabal, A., & Martin, S. (2017). General interviewer training curriculum for computer-assisted personal interviews (GIT-CAPI; Version 1, 2017). GESIS Survey Guidelines. Mannheim, Germany: GESIS – Leibniz-Institute for the Social Sciences. https://doi.org/10.15465/gesis-sg_en_022

Durand, C., Gagnon, M.-E., Doucet, C., & Lacourse, E. (2006). An inquiry into the efficacy of a complementary training session for telephone survey interviewers. Bulletin de Méthodologie Sociologique, 92(1), 5-27. https://doi.org/10.1177/075910630609200103

Eckman, S., & Kreuter, F. (2011). Confirmation bias in housing unit listings. Public Opinion Quarterly, 75(1), 139-150. https://doi.org/10.1093/poq/nfq066

ESS. (2016a). ESS interviewer briefing: NC manual. London: ESS ERIC headquarters. Retrieved from Bergen, European Social Survey Data Archive: http://www.europeansocialsurvey.org/methodology/ess_methodology/data_collection.html (2018/04/30)

ESS. (2016b). ESS round 8 interviewer briefing: Interviewer manual. London: ESS ERIC headquarters. Retrieved from Bergen, European Social Survey Data Archive: http://www.europeansocialsurvey.org/methodology/ess_methodology/data_collection.html (2018/04/30)

ESS. (2016c). Survey specification for ESS ERIC member, observer and guest countries. Retrieved from http://www.europeansocialsurvey.org/methodology/ess_methodology/survey_specifications.html (04/10/2018)

ESS. (2018). ESS-8 2016 documentation report. Edition 2.1. Retrieved from Bergen, European Social Survey Data Archive: https://www.europeansocialsurvey.org/docs/round8/survey/ESS8_data_documentation_report_e02_1.pdf (2018/05/29)

Fowler Jr, F. J., & Mangione, T. W. (1986). Reducing interviewer effects on health survey data (Report No. 141). Retrieved from Rockville, MD: National Center for Health Services Research and Health Care Technology.

Fowler Jr., F. J. (1991). Reducing interviewer-related error through interviewer training, supervision, and other means. In P. P. Biemer, R. M. Groves, L. Lyberg, N. A. Mathiowetz, & S. Sudman (Eds.), Measurement errors in surveys (pp. 259-278). Hoboken, NJ: John Wiley & Sons.

Groves, R. M. (2005). The interviewer as a source of survey measurement error. In R. M. Groves (Ed.), Survey errors and survey costs (pp. 357-406): John Wiley & Sons.

Groves, R. M., Fowler, F. J. J., Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (1992). Survey interviewing. In R. M. Groves, F. J. J. Fowler, M. P. Couper, J. M. Lep-

kowski, E. Singer, & R. Tourangeau (Eds.), Survey methodology (pp. 269-301). John Wiley & Sons.

Groves, R. M., Fowler, F. J. J., Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2009). Survey methodology (Vol. 2). Hoboken, NJ: John Wiley & Sons.

Groves, R. M., & Lyberg, L. (2010). Total survey error: Past, present and future. Public Opinion Quarterly, 74(5), 849–879. https://doi.org/10.1093/poq/nfq065

Guest, L. (1954). A new training method for opinion interviewers. Public Opinion Quarterly, 18(3), 287-299.

Hubal, R. C., & Day, R. S. (2006). Informed consent procedures: An experimental test using a virtual character in a dialog systems training application. Journal of Biomedical Informatics, 39(5), 532-540.

International Organization for Standardization (ISO). (2012). Market, opinion and social research - Vocabulary and service requirements (ISO Standard No. 20252: 2012 [en]. Geneva, Switzerland: International Organization for Standardization.

Lessler, J. T., Eyerman, J., & Wang, K. (2008). Interviewer training. In E. D. De Leeuw, J. J. Hox, & D. A. Dillman (Eds.), International handbook of survey methodology (pp. 442-478). New York, NY: Taylor & Francis Group.

Loosveldt, G. (2008). Face-to-face interviews. In E. D. De Leeuw, J. J. Hox, & D. A. Dillman (Eds.), International handbook of survey methodology (pp. 201–220). New York, NY: Taylor & Francis Group.

Malter, F., & Börsch-Supan, A. (2017). SHARE Wave 6: Panel innovations and collecting dried blood spots. Munich: Munich Center for the Economics of Aging (MEA) at the Max Planck Institute for Social Law and Social Policy (MPISOC).

Malter, F., Schuller, K., & Börsch-Supan, A. (2016). SHARE compliance profiles – Wave 6. Munich: MEA, Max Planck Institute for Social Law and Social Policy.

Mayer, T. S., & O'Brien, E. (2001). Interviewer refusal aversion training to increase survey participation. Proceedings of the Annual Meeting of the American Statistical Association, 2001.

McConaghy, M., & Carey, S. (2005). What happens after ART? Results of two experiments designed to improve response rates with interviewers at the Office of National Statistics, UK. Paper presented at the 60th Annual Conference of the American Association for Public Opinion Research, Phoenix, AZ.

Meuler, E. (2010). Didaktik der Erwachsenenbildung - Weiterbildung als offenes Projekt. In R. Tippelt & A. Von Hippel (Eds.), Handbuch Erwachsenenbildung/Weiterbildung (Vol. 4, pp. 973-988). Wiesbaden: VS Verlag für Sozialwissenschaften.

O'Brien, E. M., Mayer, T. S., Groves, R. M., & O'Neill, G. E. (2002). Interviewer training to increase survey participation. Proceedings of the 2002 Joint Statistical Meetings of the American Statistical Association.

OECD. (2011). PIAAC technical standards and guidelines. Retrieved from http://www.oecd.org/skills/piaac/documentation.htm (2014/06/12)

OECD. (2013). The survey of adult skills: Reader's companion, second edition, OECD skills studies, OECD Publishing, Paris. Retrieved from https://doi.org/10.1787/9789264258075-en (2016/12/05)

Pennell, B.-E., Harkness, J. A., Levenstein, R., & Quaglia, M. (2010). Challenges in cross-national data collection. In J. A. Harkness, M. Braun, B. Edwards, T. P. Johnson, L. E. Lyberg, P. P. Mohler, B.-E. Pennell, & T. W. Smith (Eds.), Survey methods in multi-

national, multiregional, and multicultural contexts (pp. 269-298). Hoboken, NJ: John Wiley & Sons.

Pennell, B.-E., Hibben, K. C., Lyberg, L. E., Mohler, P. P., & Worku, G. (2017). A total survey error perspective on surveys in multinational, multiregional, and multicultural contexts. In P. P. Biemer, E. D. de Leeuw, S. Eckman, B. Edwards, F. Kreuter, L. E. Lyberg, C. N. Tucker, & B. T. West (Eds.), Total Survey Error in practice (pp. 179-201). Hoboken, NJ: John Wiley & Sons.

Rutgers Online Degrees. (n.d.). The principles of adult learning theory. Retrieved from https://online.rutgers.edu/blog/principles-of-adult-learning-theory/ (2018/05/04)

Schaeffer, N. C., Dykema, J., & Maynard, D. W. (2010). Interviewers and interviewing. In P. V. Marsden & J. D. Wright (Eds.), Handbook of survey research (pp. 437–470). Bingley, UK: Emerald Group Publishing Limited.

Schnell, R., & Trappmann, M. (2006). The effect of the refusal avoidance training experiment on final disposition codes in the German ESS-2 (Working Paper 3/2006: Center for Quantitative Methods and Survey Research, University of Konstanz). Retrieved from http://nbn-resolving.de/urn:nbn:de:0168-ssoar-114189

Schouten, B., Peytchev, A., & Wagner, J. (2017). Adaptive survey design. Boca Raton, FL: CRC Press.

Smith, M. K. (2002). Malcolm Knowles, informal adult education, self-direction and andragogy. Retrieved from http://infed.org/mobi/malcolm-knowles-informal-adult-education-self-direction-and-andragogy/ (2018/05/04)

Smith, T. W. (2011). Refining the total survey error perspective. International Journal of Public Opinion Research, 23(4), 464–484. https://doi.org/10.1093/ijpor/edq052

Smith, T. W. (2019). Improving multinational, multiregional and multicultural comparability (3MC) using the total survey error (TSE) paradigm In T. P. Johnson, B.-E. Pennell, I. Stoop, & B. Dorer (Eds.), Advances in comparative surveymethods: Multinational, multiregional, and multicultural contexts (3MC) (pp. 1-21). Hoboken, NJ: John Wiley & Sons.

Survey Research Center at the University of Michigan. (2016). Guidelines for best practice in cross-cultural surveys. Survey Research Center, Institute for Social Research, University of Michigan.

Tusting, K., & Barton, D. (2003). Models of adult learning: a literature review. Retrieved from London, UK: National Research and Development Centre for Adult Literacy and Numeracy (www.nrdc.org.uk (2018/05/09)).

West, B. T., & Blom, A. G. (2017). Explaining interviewer effects: A research synthesis. Journal of Survey Statistics and Methodology, 5(2), 175–211. https://doi.org/10.1093/jssam/smw024

Zabal, A., Martin, S., Massing, N., Ackermann, D., Helmschrott, S., Barkow, I., & Rammstedt, B. (2014). PIAAC Germany 2012: Technical report. Münster: Waxmann.

Appendix

Survey Specifications and Characteristics of the ESS, PIAAC, and SHARE Across all Participating Countries

	ESS 2016 (Round 8)[1]	PIAAC (Cycle 1, Round 1)[2]	SHARE (Wave 6)[3]
Scope	Attitudes, beliefs, and behavior	Adult competencies	Health, ageing, and retirement
Survey type	Cross-sectional (first round in 2002, Round 8 in 2016) Every 2 years	Cross-sectional (first cycle in 2011/12) Multi-cycle (every 10 years)	Longitudinal (start in 2004/05, Wave 6 in 2014/15) Every 2 years
Survey mode	F2F: CAPI	F2F core questionnaire: CAPI Self-administered cognitive assessment: CASI, PAPI	F2F: CAPI Self-administered PAPI for sensitive questions in some countries[4] Physical measurements
Sampling method	Probability-based	Probability-based	Probability-based
# of participating countries	23	24	18
Target population	General, 15 years +	General, 16 - 65 years	General, 50 years +
Field period	September–December 2016	August 2011–March 2012	February–November 2015
Realized sample size per country (range)	880–2,852	3,761–27,285	1,169–6,100[5]
Interview duration (average)	60 minutes	90 minutes	80 minutes
# of trained interviewers per country (range)	41–281	70–810	---[6]

	ESS 2016 (Round 8)[1]	PIAAC (Cycle 1, Round 1)[2]	SHARE (Wave 6)[3]
Non-standard interviewer tasks[7]	No	Yes	Yes
Interviewer remuneration scheme	General approach: freelance contract, paid per completed interview	Independent of the number of completed interviews	---
Interviewer workload	Max. of 48 cases (respondents and non-respondents)	Max. of 40 completed assessments per month	On average 15 interviews; must not exceed 50
Interviewer recruitment & hiring	Responsibility of survey institute Interviewers with general training	Responsibility of survey institute Aim is to hire interviewers with at least 2 years working experience Selected from a variety of job-offering sources Applicants with various qualifications Process should start at least 8 weeks prior to the start of data collection	The appropriate number of interviewers have to be available in a sufficient regional spread

Note. Survey programs in alphabetical order; CAPI = computer-assisted personal interviewing; CASI = computer-assisted self-interviewing; ESS = European Social Survey; F2F = face to face; PAPI = paper and pencil interviewing; PIAAC = Programme for the International Assessment of Adult Competencies; SHARE = Survey of Health, Ageing and Retirement in Europe; --- = no information available in the survey documentation.

[1] ESS (2016c); Beullens, Loosveldt, Denies, and Vandenplas (2016); ESS (2018).
[2] OECD (2013).
[3] Börsch-Supan and Jürges (2005); Malter, Schuller, and Börsch-Supan (2016); SHARE (2018).
[4] Austria, Czech Republic, Greece, Israel, Slovenia, Switzerland.
[5] Girona added to Spain.
[6] 44–170 interviewers were working in the field. Information about the number of interviewers trained was not published.
[7] Refers to tasks that go beyond the standard tasks (e.g., contacting, gaining cooperation, conducting face-to-face interviews).

methods, data, analyses | Vol. 14(1), 2020, pp. 61-90 DOI: 10.12758/mda.2019.06

Solidarity and Self-Interest: Using Mixture Modeling to Learn about Social Policy Preferences

José Alemán & Dwayne Woods
Fordham University, New York & Purdue University, Indiana

Abstract

This article addresses the problem of measuring social policy preferences in a valid and reliable way. Scholars have faced a number of challenges in measuring these preferences. First, it is not clear how exactly we should conceive of this domain. Second, the literature presents contradictory findings regarding the effect of contextual factors on policy preferences. Third, abstract preferences regarding the welfare state and information about its performance can affect each other, complicating the attempt to distinguish between the two. Finally, latent manifestations of these preferences might not be equivalent across countries. We develop an approach that validly and reliably measures attitudes about the role of government in addressing inequalities in the market distribution of resources. Mixture modeling and in particular latent class analysis enables us to take advantage of information for multiple countries and survey questions while doing justice to the characteristics of the survey data. Using three waves of the International Social Survey Programme's module on social inequality, we find that preferences towards the market and the role of government in the economy form four distinct clusters of individuals that we refer to as "moderate altruists", "moderate egoists", "extreme altruists", and "extreme egoists". These clusters tend to be homogenous with respect to both abstract notions of the role the government should play in the economy as well as about evaluations of actual performance. The exceptions are the last two survey waves, for which we find that one class exhibits a mixed profile of individuals: solidaristic with respect to some indicators, but self-interested with respect to others.

Keywords: solidarism, self interest, social policy preferences, latent class analysis, mixture modelling

A voluminous social science literature treats solidarism, or care about the well-being of others, as "unpredictable 'social noise'" (Dimick et al. 2018, p. 442). Our study explores how to conceive of this pre-disposition (Cavaillé & Trump 2015; Dimick et al. 2017; Fong 2001). We do this using a latent class modeling framework that considers not only individual and country level determinants of these preferences, but also the equivalence of latent constructs across countries. Our research builds on recent work using categorical variables to capture latent preferences, and provides an approach to deal with lack of independence among some indicators used to represent them. In so doing, we can reveal preferences in a valid and reliable way.

Using three waves of the International Social Survey Programme's module on social inequality, we find that preferences towards the market and the role of government in the economy form four distinct clusters of individuals that we refer to as "moderate altruists", "moderate egoists", "extreme altruists", and "extreme egoists". These clusters tend to be homogenous with respect to both abstract notions of the role the government should play in the economy as well as about evaluations of actual performance. The exceptions are the last two survey waves, for which we find that one class exhibits a mixed profile of individuals: solidaristic with respect to some indicators, but self-interested with respect to others.

The following section discusses the challenges inherent in accurately measuring social policy preferences. In section two, we introduce latent class analysis (a form of mixture modeling) and discuss its advantages over alternatives. We then apply this methodology to the task of revealing preferences in cross-national surveys. Section five examines how robust our results are to alternative classifications. We conclude with some observations for future research.

Measuring Preferences in Survey Research: Empirical Challenges

Scholars have faced a number of challenges in measuring social policy preferences. It is not clear, for example, how exactly we should conceive of this domain. Arts

Acknowledgments

We would like to thank two anonymous reviewers, Ida Bastiaens, Petra Brien, Eldad Davidov, Noam Gidron, Alex Kustov, Vincent Mahler, Guy B.D. Moors, Melissa Patel, Daniel Treisman, participants at the APSA annual meeting panel on inequality and especially Nate Breznau for their help with various aspects of researching and writing this paper.

Direct correspondence to

José Alemán, Professor of Political Science, Fordham University, Bronx, New York
aleman@fordham.edu

and Gelissen (2001) report that attitudes towards "solidaristic" policies cluster in one dimension, implying that individuals either support these policies or oppose them.[1] Alesina and Angeletos (2005) conceive of self-interest and solidarity as a variable that ranges from identifying most closely with a libertarian ideal of markets as natural, efficient, and fair, to believing that markets not always work this way and should not be the sole criterion used to make allocative decisions.

Other work implies that "other-regarding preferences" (Dimick et al. 2018) are not one-dimensional. Jensen and Petersen (2017, p. 68) claim for example that individuals see recipients of health care as deserving compared to recipients of unemployment compensation. Cavaillé and Trump (2015) similarly claim that redistribution can take on two meanings – taking from the rich and giving to the poor. Finally, Rehm et al. (2012, p. 390) find that when asked to evaluate social programs in the abstract, individuals tend to favor them due to loss aversion, the tendency to weigh potential losses in benefits more than potential increases in one's post-tax income.[2]

Scholars are also unsure what effects contextual factors have on policy preferences. Dimick et al. (2017, p. 386) find that "an increase in macro-inequality will lead to more support for redistribution", particularly among the rich.[3] Conversely, Kelly and Enns (2010) and Trump (2018) find that it reduces support (irrespective of income) for this policy (Cavaillé & Trump 2015, p. 157). These findings, however, are based either on experimental data from a few counties or on longitudinal evidence from the United States. Two studies with wide country-year coverage find no effect of country-level inequality on support for redistribution (Breznau and Hommerich 2019; VanHeuvelen 2017).

A final set of challenges concerns the potential for perceptions of how the welfare state is performing to prime abstract preferences about the desirability of social policies. As Trump (2018) notes, perceptions of inequality strongly predict whether individuals see income differences as legitimate. Gimpelson and Treisman (2018, p. 30) cite Niehues (2014) to the effect that "a correlation between perceived inequality and the belief that it" is "too high", as well as between perceived inequality and preferences for redistribution", exists.[4] More specifically, "perceived

1 In social science research, solidarity is defined as concern for one's group (Dimick et al. 2017, p. 387), whether the group is one's class, ethnicity, or nation. Following the literature, we see support for policies such as income redistribution as evidence of social solidarity because these policies can benefit others in addition to oneself or others at one's expense. Below, we also evaluate whether this relationship depends on one's personal income.

2 In Kahneman and Tversky's (1979, p. 279) words, "losses loom larger than gains" in people's minds.

3 See also Schmidt-Catran (2016).

4 Niehues derived these correlations using the same ISSP data for Wave IV that we use here. See Kim et al. (2018) also.

inequality rather than actual inequality significantly affects redistributive preferences" (Choi 2019, 4). The opposite, preferences affecting perceptions, also occurs, as "more anti-redistributive preferences predict believing taxes on high earners are too high." (Gingrich 2014, p. 578).

We need a methodological approach then that empirically allows for the possibility that abstract preferences regarding the welfare state and information about its performance can influence each other simultaneously. This would help us move beyond the current impasse in the literature between standard accounts favoring self-interest and more recent works that also expect individuals to care about others. Before this can happen, however, we need to put solidaristic attitudes on solid empirical ground.

The Latent Class Approach

Scholars study social policy preferences using either a single survey prompt or a latent variable framework. In the latter case, they typically rely on principal component or a similar factor analytic technique. Latent class analysis allows for more flexibility because "there is no need for normality assumptions as there is in factor analysis": "instead of assuming that" [indicator] "variables follow any particular distribution within the classes", "LCA lets the variables follow any distribution, as long as they are unrelated to each other (independent) within classes". (Oberski 2016, p. 7).

The latent class approach is especially useful given recent work demonstrating that the multidimensionality of welfare state attitudes cannot be adequately captured using only linear measurement models (Kulin et al. 2016; Roosma et al. 2013). As these works make clear, individuals vary not only in their preferences regarding what welfare states do, but also in their preferences about what welfare states should do.[5] This is because people are able to distinguish "the welfare state's goals and range" from "it's efficiency, effectiveness, and policy outcomes". (Roosma et al. 2013, p. 235). Accordingly, they could strongly favor the welfare state both con-

5 Individuals, in other words, vary on "the should and is aspects of welfare attitudes" (Roosma et al. 2014, p. 201). This is because "the public has both a relative preference for policy and an absolute preference" (Soroka and Wlezein 2010, p. 25). We don't necessarily see the relationship between the two mechanically, however, as Soroka and Wlezein's "thermostatic model" implies. In this model, the public's relative preference represents the difference between its "preferred level of policy…and the level it actually gets". In reality, individuals rely on heuristic shortcuts to form their views, particularly when demands on their cognitive capacity are high. They thus display what is known as "bounded rationality" (O'Grady 2017). This explains why "preferences for redistribution and social spending", once formed, only change in response to "large changes in economic circumstances" (O'Grady 2017).

cretely and in the abstract, oppose it on both grounds, embrace an ambitious role for social policy in the abstract while being critical about its outcomes, or approve of outcomes while being critical of stated goals. The four possible attitudinal profiles, moreover, can manifest themselves differently across countries.

If, as alluded to above, individuals' perceptions of how the welfare state performs affect their attitudes about what the welfare state should do and vice-versa, we need a methodology that can handle these "possible feedback effects" (Roosma et al. 2014, p. 201).[6] In latent class modeling, interactions between the latent variable and indicator variables, usually omitted, enables consideration of these effects. As noted above, it is usually assumed that the observed indicators are mutually independent (or uncorrelated) conditional on the latent variable (Oberski 2016, p. 11). This requires the omission of all interaction terms between the latent construct and indicator variables (hence the independence assumption). Relaxing this assumption enables consideration of feedback effects by specifying higher order interaction terms. (Magidson & Vermunt 2001, p. 226).

The model essentially asks how likely a subject is to belong to one of N categories in a nominal variable we dub *solidarity*. Individuals are then grouped into exclusive subpopulations "based on similar patterns of observed cross-sectional and/or longitudinal data." (Berlin et al. 2014, p. 175). The resulting classes are "characterized not by exact response patterns but by response *profiles* or typologies described by the relative frequencies of item endorsements" (Masyn 2013, p. 556). Predictor variables can be used to facilitate the placement of observations into classes, in which case the goal is to examine whether covariates can explain "mean differences in outcomes across latent classes" (Berlin et al. 2014, p. 175).

Studying multiple policies and countries can pose problems if latent constructs are non-invariant cross-nationally (Alemán & Woods 2016). To avoid problems with measurement invariance, researchers typically rely on dichotomized versions of survey indicators (VanHeuvelen 2017, p. 49). One advantage of LCA, which has not been widely used to explain social policy preferences, is that it provides a rigorous and systematic framework for investigating construct equivalence (Kankaraš & Moors 2009; Moors 2004). The approach, dubbed "multigroup latent class structure modeling", can easily diagnose and accommodate several forms of parameter heterogeneity.7

In sum, a latent class approach allows us to measure a construct that cannot be perfectly measured while doing justice to the data generating process (Oberski

6 Roosma et al. describe these different dimensions, but not their possible feedback effects.

7 Similar approaches such as multigroup confirmatory factor analysis (MCFA) exist for models with continuous indicator and latent variables. Multigroup latent class structure modeling, however, outperforms its counterparts (Kankaraš et al. 2011).

2016). We believe this method elicits preferences about social policy based on individual characteristics and exposure to varying contexts.

Data Sources and Variables

We use public opinion data from the International Social Survey Programme (hereafter ISSP) to examine whether individuals can be sorted into classes based on their attitudes towards the market allocation of resources and the role of government in molding this allocation. One advantage of the ISSP is that it has carried out periodic surveys of attitudes towards social inequality (the Social Inequality series). These questionnaires, administered in 1987, 1992, 1999, and 2009, target a variety of countries, mostly democracies. We are able to use all survey waves except the third one, which did not provide enough information to standardize the income or earnings of survey respondents. Despite the varying number of countries, years, and individuals surveyed, our goal is to find similarities in this heterogeneity.[8]

Table 1 presents a list of questions that can be used to assess social policy preferences, along with the year(s) the survey wave containing the question was administered. Our choice of questions was motivated by our desire to tap into preferences regarding the goals and capabilities of the welfare state, as well as to evoke assessments of government efforts in targeting particular groups (i.e., the unemployed, the poor, students, the middle class). One advantage of the ISSP is that all questions have the same ordinal ranking, with 1 usually implying strong agreement, 3 neutrality, and 5 strong disagreement. To facilitate analysis and interpretation, we recoded some variables so as to have higher values denote increasing social solidarity or progressivism.[9]

While there is much continuity in questions from survey to survey, some questions are missing from some of the waves.[10] We consider this an advantage since we are trying to estimate attitudes that are latent and as such, do not exhibit a perfect correspondence with our survey instruments.

Of the twelve questions displayed in Table 1, some clearly elicit general beliefs about the fairness of the market mechanism and the role that government plays in shaping it, while others evoke an evaluation of the status quo. We first selected

8 The number of countries in the analysis, which is based on data availability, ranges from five in 1992 to thirty-one in 2009. Appendix A contains a list of countries we studied, organized by wave.

9 Following standard practice, we excluded from the analysis respondents who are unsure or uncooperative.

10 We were able to use most questions fielded, except for those which contained more missing than complete observations – poor and unemployed in Wave II and university in Waves II, III, and IV. The percent of missing observations for unemployed in Wave II, for example, is 88.67, while for poor it is 88.77.

Table 1 Indicators of solidarity/self-interest

Question	Variable name	Years asked
It is the responsibility of the government to reduce differences in income between people with high incomes and people with low incomes[11]	government responsibility	1987, 1992, 2009
The government should provide a decent standard of living for the unemployed	unemployed	1987, 2009
The government should provide more chances for children from poor families to go to university	university	1987
The government should provide a job for everyone who wants one	job guarantee	1987, 1992
The government should provide everyone with a guaranteed basic income	basic income	1987, 1992
Is it just or unjust - right or wrong - that people with higher incomes can buy better health care than people with lower incomes?	private health care just	2009
Is it just or unjust - right or wrong - that people with higher incomes can buy better education for their children than people with lower incomes?	private education just	2009
The government should spend less on benefits for the poor	poor	1987, 2009
Differences in income in [respondent's country] are too large	income differences	1987, 1992, 2009
Generally, how would you describe taxes in [respondent's country] today for those with high incomes?	top taxes	1987, 1992, 2009
Do you think people with high incomes should pay a larger share of their income in taxes than those with low incomes?	progressive taxation	1987, 1992, 2009
Inequality continues to exist because it benefits the rich and powerful	inequality helps the rich	1987, 1992

questions that we thought tap abstract attitudes and perceptions, moving then to those that seem to elicit a comment on the status quo. The first seven questions evoke abstract beliefs about economic fairness while questions eight through ten are evaluative. Based solely on their phrasing, question eleven seems to probe

abstract attitudes towards inequality and redistribution, while question twelve lends itself to both kinds of interpretation.

Existing studies provide a mixed picture regarding the effects of demographic variables on social policy attitudes (Breznau 2010, p. 476). We control for these characteristics since they are standard in the public opinion literature. We also control for several country-level variables that have featured prominently in the literature.

Sex is a dichotomous variable taking the value of 1 for females and 0 for males. In the literature, men are generally shown to exhibit less solidarity than women.

Age ranges vary by survey wave but for the population as a whole it is a continuous variable ranging in value from 15 to 98.

Education. Competition from immigrants may cause workers with little education to oppose programs that could be construed as enhancing the labor market prospects of other similarly skilled workers (Alt & Iversen 2017, p. 21; Kunovich 2009, p. 575). An additional factor bearing on the preferences of dissimilarly educated workers is the extent to which education increases class solidarity. As Kunovich (2009, p. 575) notes, "[i]ndividuals with greater cognitive skills (i.e., more education) … can better imagine belonging to larger groups". This implies a positive correlation between education and solidaristic attitudes.

The literature on the link between labor market risks and welfare state attitudes, however, makes a convincing case that better-educated individuals have more skills, which could imply that they have more stable income streams, anticipate upward mobility more, and need social policies less (O'Grady 2017, p. 5). This raises the possibility that education increases self-interest and vice versa (Alesina & Giuliano 2011, p. 21; Breznau 2010, p. 461; Gimpelson & Treisman 2017, p. 19).

In Wave I, education is a categorical variable with nine categories ranging from "None, still at school" to "Complete University", with adjustments in the number of categories made for certain countries reflecting variation in educational systems around the world. In Wave II, education refers to years of schooling, which is a continuous variable. In Wave IV, education is a categorical variable with 'no formal qualification' as the first category followed by 2) lowest formal qualification; 3) above lowest qualification; 4) higher secondary level completed; 5) above higher secondary level; and 6) university degree completed.

Personal income. The median voter theory (Meltzer & Richard 1981), the bedrock of much political economy work, predicts a negative relationship between pretax and -transfer income and demand for redistribution. We thus expect that "the (relatively) poor support redistribution more than the (relatively) rich" (Dimick et al. 2017, p. 386).

---

11 According to Choi (2019, p. 15), this is "the most widely used measure of redistributive preferences in empirical studies."

The ISSP provides two measures of personal well-being, one labeled "family income" and the other "earnings". For some countries the measures refer to pre-tax and -transfer earnings and for others to net income. Whether individuals correctly perceive their income as being pre-tax and -transfer or net is questionable, but this is not likely to bias the results unless these perceptions are non-randomly distributed. In addition, in some countries individuals were asked to report monthly, in others yearly amounts. Finally, the precise amounts reported by survey participants in Waves I and II contains a lot of missing data.

We could use self-reported social class in lieu of a more objective measure of welfare. Subjective measures, however, "also capture psychological elements besides actual income" (Midtbø 2017, p. 6). This poses a problem if the two vary greatly or in ways that are unknown across countries. In all three survey waves we study, moreover, earnings and family income are moderately correlated, while subjective social class correlates weakly with both. A measure in Waves I and II that provides income and earning brackets for respondents to choose from is more complete. For these waves, we thus follow Dimick et al. (2018) in creating two variables using the robust Pareto midpoint estimator (von Hippel et al. 2016). These variables contain the midpoint yearly income and earnings corresponding to each reported category, "while the value for the final open-ended bin is imputed from a Pareto distribution" (Dimick et al. 2018, p. 452). Since the amounts reported are in local currencies, we calculated standard deviations from the country mean and used those in our models (Dion & Birchfield 2010, pp. 321-322; Rehm 2011, p. 279). For Wave IV, we are able to use the income and earnings figures individuals reported.[12]

Redistribution. Spending (of tax receipts) by governments on social programs accounts for much of the variation across democracies in "redistributive effort". "Spending questions […however,] ask people about priorities relative to very different national baselines." (Rehm 2012, p. 399). What is needed then is a measure of relative redistribution, or absolute redistribution divided by market inequality. Our measure of income redistribution is thus the reduction in the Gini coefficient due to taxes and transfers as a ratio of this coefficient (Solt 2016).

GDP per capita. Individuals in less developed and highly unequal societies seem more concerned with the needs of others than their counterparts in more developed and egalitarian societies (Dion & Birchfield 2010; VanHeuvelen 2017). Fong (211, p. 242) similarly claims that perceived poverty increases support for redistribution among high-income earners. To assess the effect of development on attitudes towards social policy, we use a measure of real GDP in 2011 US dollars given in purchasing power parity (or PPP) terms. We divide this measure by a country's population to obtain per capita measures. The Penn World Table (Feenstra et al. 2015) is the source for these variables.

12 One benefit of having income and earnings data in local currencies is that this method of accounting minimizes errors.

Economic growth. Economic growth could facilitate solidaristic tendencies by making people better off. If a majority believe, however, "in insuring industrious people against bad luck, but not providing unconditional assistance to the poor if their condition is due to idleness" (Fong 2001, p. 242), individuals may be less likely to care for others when they regard the economic environment as good. We represent economic growth using a measure of inflation-adjusted growth in GDP per capita from the World Development Indicators (World Bank 2017).

Unemployment rate. Some have claimed that unemployment should increase support for welfare policies (Breznau 2010, p. 13; VanHeuvelen 2017, p. 45). Wehl (2018) however finds that unemployment does not significantly explain support for labor market policies. VanHeuvelen (2017) also found that unemployment does not significantly increase support for redistribution. Our variable refers to those who are unemployed in a given year as a share of the active labor force. This data, originally compiled by the International Labor Organization, was similarly derived from the WDI dataset.

Employment status. An important question is whether employed and unemployed respondents regard social policy in a similar fashion. Some have claimed that the employed, also known as insiders in countries with labor market dualities and high unemployment, favor government programs that insure or redistribute income if the beneficiaries are insiders like themselves and not the unemployed (Moene & Wallerstein 2001, 2003; Rueda 2007).

Church attendance. An important literature has claimed that religiosity makes individuals disapprove of social insurance even when they stand to benefit from it (De la O & Rodden 2008; Scheve & Stasavage 2006). Poor religious voters accordingly prioritize moral issues. This could make these individuals appear less solidaristic than secular ones.[13] Breznau (2010, p. 474) found, however, that church attendance had "little to no influence on [welfare] policy preferences". We evaluate these expectations using a question about the frequency of attending religious services. For Waves I and II, we use a categorical variable with six categories, whereas in Wave IV the same variable contains eight categories.

Partisanship. A large literature has claimed that "Left-Right placement bundles together a variety of policy attitudes and value orientations … the strongest of which are attitudes connected to the extent of state involvement in the economy and the limits to redistribution" (Bosancianu 2017, p. 1592). We thus include in our models a measure of partisan affiliation that ranges from far-left to far-right and also includes choices for "Other, no specific party" and "No party preference".

13 Aversion to social insurance, however, should not be taken to imply that religious individuals cannot behave altruistically by, for example, donating money to their churches or other charities. Logically speaking, these individuals could be very altruistic in the private sphere, while opposing government social programs on principle and/or based on their performance.

Before reporting our findings, we note that for Waves I and IV, values are given for 1986 and 2008 respectively. For Wave II, because the year of fieldwork was in some cases 1993 and in one case 1991, values given for the variables are for 1992. Regarding our specifications, due to the ordered and categorical nature of our indicators, we use ordered logistic regression for the measurement portion of the model. The probability of placing in one of the classes is modeled using multinomial logistic regression. We use sampling weights to account for over- and under-sampled observations.[14]

Exploratory Analysis

We begin by noting that we follow a specific model development strategy before settling on our preferred specification (Vermunt & Magidson 2005, p. 43). First, we estimate unconditional models (or models without covariates) with 2, 3, and 4 latent clusters. We then add covariates to these models (conditional estimation) to improve model fit. At every step, we examine the log likelihood (LL) and the Bayesian Information Criterion (BIC) for information on parsimony and fit, respectively. Generally speaking, lower values for these statistics indicate a better fit.

For all three waves, a model with 4 clusters fits the data best, as evidenced especially by the BIC. Adding covariates also led to large reductions in this statistic, confirming their role in helping to measure the latent variable. For all waves, we also explored construct equivalence. As Nagelkerke et al. (2016) point out, the assumption of unit independence is automatically violated when observations are nested in groups, as in many studies featuring surveys conducted in multiple countries. In this case, it is important "to detect misfit that originates from the model not fitting particular groups as well as others." (Nagelkerke et al. 2016, p. 255). Nagelkerke et al. define a between-group bivariate residual that is calculated by using the grouping variable as a nominal covariate with its effect set equal to 0 (Vermunt & Magidson 2016, p. 121). The model is then estimated and residuals examined between pairs of indicator variables, pairs of covariates and indicators, and between the grouping variable and indicators. The latter in particular can be evaluated for evidence of parameter heterogeneity across countries.

> "[L]arge residuals indicate large direct effects of particular group variables…If…large residuals are associated with group variables, an appropriate strategy is to include the direct effects of the group variable with the largest residuals, re-estimate the model and check the updated residuals after this new model is estimated. This procedure can be repeated until all

14 The variables used for weighting are *v107*, *v176*, and weight respectively.

> of the residuals between group variables and response variables become small." (Moors 2004, p. 309).

In Wave I, all ten bivariate residuals between the grouping variable and indicators exceeded 3.84.[15] In Wave II's case, 4 out of 7 bivariate residuals exceeded this value. For Wave IV, the number is 8 out of 8. We thus concluded that there was significant country-level heterogeneity in parameters in all three specifications. Consequently, we proceeded to explore the possibility of modeling this heterogeneity using various kinds of random effects/multilevel models (Henry & Muthén 2010; Vermunt 2003). Once again, a multilevel model can be compared to a model without random effects using the LL and BIC test statistics.

The simplest random effects specification is a parametric model in which intercepts for the latent classes are allowed to vary randomly. In these models, the individual-level latent classes vary in size by country, but all other parameters are "fixed". The random effects themselves follow a continuous distribution of means across groups. There is also a non-parametric version of this model which conceives of countries not as continuous random means, but as belonging to a smaller set of discrete groups that in turn affect the intercepts of the individual classes (Henry & Muthén 2010). The grouping of countries in the varieties of capitalism literature offers an example of the ways in which countries could be modeled in the level 2 analysis (as liberal or coordinated market economies) (e.g., Larsen 2008). In this case, at least two country-level classes need to be specified. Figure 1 provides a visual summary of our 4-cluster solution for the indicator variables in Wave I.

As Figure 1 indicates, four classes are clearly delineated in the ten indicator variables used to measure attitudes towards social policy in Wave I. The cluster comprising the most members is Cluster 1, which appears to be composed of individuals who are moderately solidaristic. The second largest cluster is reserved for moderately self-interested individuals, with "extreme egoists" and "extreme altruists" a distant third and fourth places respectively. This plot confirms what scholars have recently observed, that the welfare state in advanced capitalist democracies is popular (Roosma et al. 2013) and that this is in part due to loss aversion (Rehm et al. 2012). The best fitting model for Wave I, it turns out, is a non-parametric estimation with two country-level latent classes affecting the intercepts for the individual level classes. It is not hard to see how this would occur in a model where *education* does not have a uniform number of categories across countries.

It is important to note that for the tables that follow, a positive coefficient implies that a particular variable is more likely to place/keep individuals in a certain class, whereas a negative one indicates that the variable is likely to place individuals in a different class. Table 2 presents the results for Wave I.

---

15 "For 1 degree of freedom effects, bivariate residuals larger than 3.84 indicate statistical significance at the .05 level. (Vermunt & Magidson 2005, p. 125).

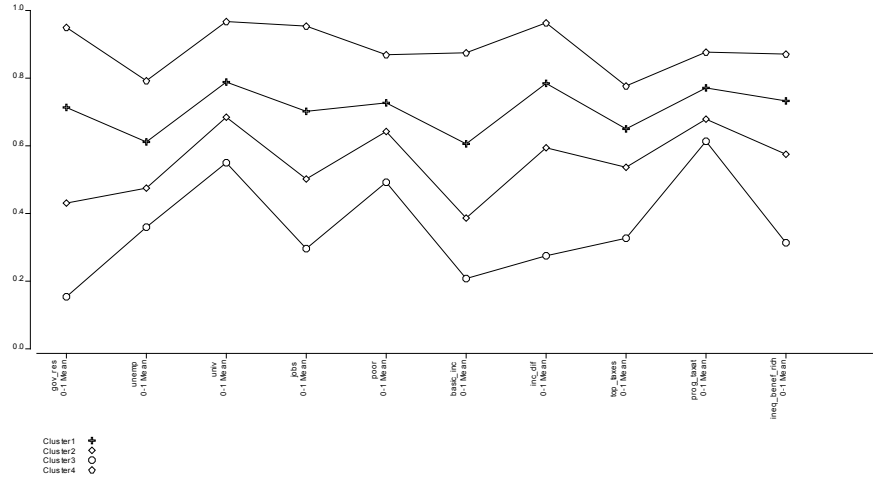


Figure 1 Profile plot of cluster solution for the latent class analysis of Wave I

As Table 2 indicates, specifying random effects is warranted – level-2 classes affect intercepts for individual-level clusters in a statistically significant way. In our discussion of these individual-level results, we speak primarily about Cluster 1, the largest class of individuals. As Table 2 indicates, all covariates significantly explain placement into a particular class. As expected, *females*, those with little education, the *unemployed*, those who lean left ideologically, and the less well-off tend to be more solidaristic than *males*, those who are more educated, the *employed*, the better off economically, and right of center individuals. Regarding the country level variables, *economic growth* and *redistribution* are negatively associated with social solidarity while *unemployment* has a positive association. Contrary to claims made recently regarding the effect of religiosity on preferences towards social policy (De la O & Rodden 2008; Scheve & Stasavage 2006), we find that more religious individuals exhibit more social solidarity than less religious ones. Finally, the *age* of the respondent is not predictably associated with a particular orientation across clusters.

Table 2 also provides a model for the indicators with an R^2 that captures how well the latent variable explains these. There is evidence that the latent variable is primarily picking up attitudes about income differences and what role the government should have, if any, in reducing them because *government responsibility* and *income differences* have the highest R^2s. *Government responsibility* elicits abstract preferences or attitudes about the welfare state, while *income differences* is a comment on the status quo.

Similar to Wave II, a model with four clusters is more parsimonious and fits the data best in the case of Wave II. This time, however, the addition of random parameters does not bring about an improvement over our baseline model. As a result, we retain a model with 4 clusters whose main difference with respect to Wave I is that there is now a group of individuals (Cluster 4) who exhibit a mixed attitudinal profile: they are rather self-interested in their conception of what the welfare state should do (reduce income differences and guarantee everyone a job and a basic income), but progressive in their evaluation of its results. Once again, "moderate altruists" lead in numbers, but "moderate egoists" do not make up the second most numerous class. Instead, Cluster 2 is composed of individuals who are very self-interested, followed by individuals who are very solidaristic (Cluster 3). Figure 2 provides a visual summary of the solution for Wave II.

Table 2 Multilevel LCA of attitudes towards social policy in seven countries (1987)

Model for Indicators	Wald	p-value	R^2
government responsibility	208.971	0.000	0.584
unemployed	17.147	0.001	0.200
university	187.976	0.000	0.243
jobs	195.038	0.000	0.379
poor	143.626	0.000	0.145
basic income	53.053	0.000	0.367
income differences	168.399	0.000	0.540
top taxes	50.525	0.000	0.189
progressive taxation	123.203	0.000	0.164
inequality benefits the rich	87.934	0.000	0.307

Model for Clusters Intercept	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Wald	p-value
N	1404	1318	450	321		
group class 1	3.009	0.651	-3.744	0.084	163.092	0.000
group class 2	2.351	1.086	-2.474	-0.963	144.747	0.000
Covariates						
sex						
male	-0.138	-0.059	0.311	-0.114	148.023	0.000
female	0.138	0.059	-0.311	0.114		
age	-0.008	-0.007	0.009	0.006	23.450	0.000

education						
none/still at school	0.816	-0.061	-1.432	0.677	109717.219	0.000
	0.579	0.030	-1.786	1.177		
	0.054	-0.067	-0.348	0.361		
	-0.060	0.125	0.162	-0.226		
	-0.248	0.068	0.433	-0.254		
	-0.127	0.156	0.319	-0.348		
	-0.413	0.042	0.772	-0.401		
	-0.329	-0.170	1.099	-0.600		
complete university	-0.273	-0.123	0.781	-0.385		
Model for Clusters						
Intercept	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Wald	p-value
GDP growth	-0.152	-0.539	-0.813	1.504	120.001	0.000
unemployment	0.011	0.084	0.161	-0.255	143.224	0.000
employment status						
unemployed	0.068	-0.025	-0.518	0.475	85.964	0.000
employed	-0.068	0.025	0.518	-0.475		
income	-0.077	0.121	0.320	-0.363	33.628	0.000
redistribution	-0.012	-0.021	0.001	0.032	52.767	0.000
GDP	0.000	0.000	0.000	0.000	28.592	0.000
partisanship						
far left	4.044	-3.865	-4.311	4.131	110428.720	0.000
left	1.094	0.412	-1.147	-0.359		
center	0.313	0.786	0.523	-1.621		
right	-0.032	0.763	1.256	-1.987		
far right	-6.586	1.949	3.586	1.051		
other, not specified	0.827	-0.392	0.095	-0.530		
no party preference	0.340	0.347	-0.003	-0.684		
church attendance						
once a week	0.114	0.003	-0.376	0.259	41109.523	0.000
1-3 times a month	0.247	0.167	-0.397	-0.017		
several times a year	-0.066	-0.096	0.504	-0.342		
once or twice a year	-0.183	0.088	-0.067	0.163		
less frequently	-0.051	0.027	0.375	-0.351		
never	-0.061	-0.190	-0.039	0.290		
Model for group classes						
Intercept	Class 1	Class 2	Wald	p-value		
	0.5152	-0.5152	1.584	0.21		
Overall N	3345.32					

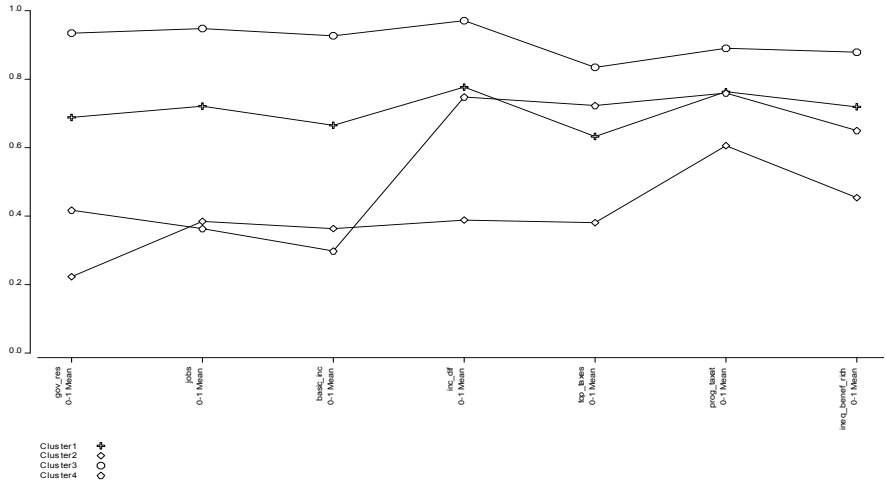


Figure 2 Profile plot of cluster solution for the latent class analysis of Wave II

Other differences between Waves 1 and 2, albeit minor, are that in the latter case an individual's employment status does not emerge as a statistically significant predictor of his/her attitudes about social policy. In addition, both *GDP growth* and personal *income* are associated with moderate social solidarity (they both increase the likelihood of placing in Cluster 1). Table 3 presents the results for this model.

Table 3 Multilevel LCA of attitudes towards social policy in five countries (1992)

Model for Indicators	Wald	p-value	R^2
government responsibility	346.311	0.000	0.559
jobs	402.222	0.000	0.462
basic income	444.764	0.000	0.433
income differences	617.964	0.000	0.498
top taxes	401.919	0.000	0.262
progressive taxation	358.124	0.000	0.208
inequality benefits the rich	321.064	0.000	0.234

Model for Clusters Intercept	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Wald	p-value
N	2872	921	808	753		
	10.567	11.894	-7.418	-15.044	71.550	0.000
Covariates						
sex						
male	-0.125	0.183	-0.163	0.105	37.134	0.000
female	0.125	-0.183	0.163	-0.105		
age	-0.010	-0.005	0.000	0.014	22.072	0.000
education	-0.073	0.122	-0.100	0.051	105.759	0.000
GDP growth	0.834	1.088	0.498	-2.419	107.419	0.000
unemployment	0.163	0.420	0.068	-0.651	30.871	0.000
employment status						
unemployed	0.050	-0.005	0.033	-0.079	3.124	0.370
employed	-0.050	0.005	-0.033	0.079		
income	0.157	0.475	-1.023	0.391	48.010	0.000
redistribution	-0.053	-0.164	0.146	0.071	71.414	0.000
GDP	0.000	0.000	0.000	0.001	146.580	0.000
partisanship						
far left	1.495	-4.760	2.228	1.037	309.598	0.000
left	0.291	0.142	-0.093	-0.340		
center	-0.078	0.849	-0.598	-0.173		
right	-0.553	1.774	-1.028	-0.193		
far right	-0.125	0.641	-0.450	-0.066		
other, not specified	-0.783	0.629	0.261	-0.107		
no party preference	-0.248	0.727	-0.321	-0.158		
church attendance						
once a week	0.101	-0.225	-0.144	0.268	55.588	0.000
1-3 times a month	0.165	-0.244	0.059	0.019		
several times a year	0.049	0.210	-0.176	-0.084		
once or twice a year	-0.032	0.155	-0.151	0.029		
less frequently	-0.138	0.083	-0.032	0.087		
never	-0.145	0.021	0.443	-0.319		
Overall N	5354					

We turn now to Wave IV, which also yields four clusters. Figure 3 provides a visual summary of this solution.

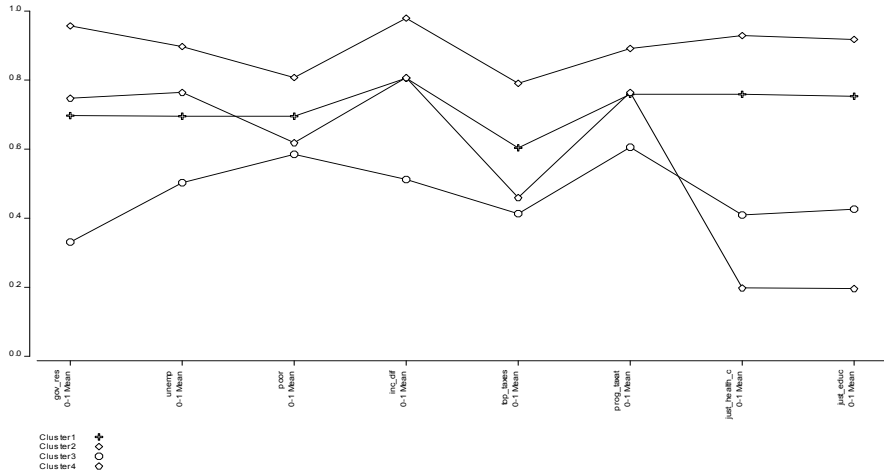


Figure 3 Profile plot of cluster solution for the latent class analysis of Wave IV

Figure 3 indicates that once again, the most numerous class is composed of individuals who are moderately solidaristic (Cluster 1). As with Wave II, there is also a class of individuals that has a mixed profile of attitudes, and together they make up the second largest group (Cluster 4). The third largest group is composed of individuals who are moderately self-interested (Cluster 3). The least numerous class (Cluster 2) groups individuals who are very solidaristic. Table 4 presents full results for this model.

The most notable differences that emerged with respect to previous results are as follows. First, the *unemployment* rate is now associated with a significant decrease and *redistribution* with a significant increase in solidaristic attitudes. Second, being *employed* is now associated with a positive and being *unemployed* with a negative propensity for moderate solidarism, although these coefficients are not highly significant statistically. Third, far-left partisanship and attending religious services several times per week are negatively associated with moderate solidarity, although the association of far-left partisanship with extreme solidarity is positive. Fourth, the indicators that are best explained by the latent variable are the ones unique to this wave asking how just it is that people with higher incomes can buy better health care and education than people with more modest means. Finally, we found that the Wald test statistic and its associated p-value cannot be computed for *GDP per capita*.

Table 4 LCA of attitudes towards social policy in thirty-one countries (2009)

Model for Indicators	Wald	p-value	R^2			
government responsibility	1651.472	0.000	0.449			
unemployed	1041.153	0.000	0.226			
poor	568.812	0.000	0.064			
income differences	1351.697	0.000	0.351			
top taxes	1229.374	0.000	0.220			
progressive taxation	1463.533	0.000	0.207			
private health care just	1625.358	0.000	0.605			
private education just	1567.072	0.000	0.581			
Model for Clusters						
Intercept	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Wald	p-value
	0.095	-0.839	-0.412	1.157	93.127	0.000
N	14637	2698	2718	3087		
Covariates						
sex						
male	-0.088	-0.114	0.156	0.046	106.709	0.000
female	0.088	0.114	-0.156	-0.046		
age	-0.001	0.010	-0.007	-0.002	77.039	0.000
education						
no formal qualification	0.048	-0.279	0.100	0.131	167.936	0.000
lowest formal qualification	0.041	0.093	-0.278	0.144		
above lowest qualification	0.152	0.272	-0.482	0.058		
higher secondary completed	-0.011	0.046	-0.005	-0.031		
above higher secondary level, other qualification	-0.080	-0.011	0.297	-0.207		
university degree completed	-0.151	-0.122	0.368	-0.095		
GDP growth	0.061	0.000	-0.009	-0.051	65.469	0.000
unemployment	-0.004	-0.027	-0.019	0.049	66.280	0.000
employment status						
unemployed	-0.001	0.015	-0.072	0.058	10.479	0.015
employed	0.001	-0.015	0.072	-0.058		
income	-0.025	-0.277	0.279	0.022	186.518	0.000
redistribution	0.003	0.048	-0.019	-0.032	646.900	0.000
GDP	0.000	0.000	0.000	0.000	.	.

partisanship						
far left	-0.036	0.900	-0.760	-0.105	909.720	0.000
left	0.156	0.449	-0.701	0.095		
center	0.059	-0.170	0.113	-0.002		
right	-0.092	-0.771	0.937	-0.074		
far right	-0.058	-0.398	0.429	0.026		
other, not specified	-0.008	0.104	-0.175	0.079		
no party preference	-0.022	-0.114	0.156	-0.019		
church attendance						
several times per week	-0.194	-0.173	0.161	0.206	208.831	0.000
once a week	0.032	-0.214	0.166	0.015		
2 or 3 times a month	-0.033	-0.160	0.286	-0.093		
Once a month	-0.040	-0.158	0.051	0.147		
Several times a year	0.051	0.277	-0.338	0.010		
Once a year	0.048	0.068	-0.155	0.039		
less than once a year	0.117	-0.024	0.074	-0.166		
never	0.019	0.383	-0.245	-0.157		
Overall N	23,426					

Robustness Checks

We look for possible deviations from the assumption that indicator variables are conditionally independent (that is, unrelated to each other within classes) and respecify our models. In so doing, we retain the most parsimonious model possible (i.e., the one with the smallest number of additional parameters), while improving model fit.

Conditional independence can be examined by looking at the correlation of indicator variables by class both before and after observations have been grouped into classes. We found that some of the indicator variables in Waves I and II had moderately significant correlations prior to observations being sorted into classes. After being sorted into classes, however, these pairwise correlations became statistically insignificant and/or very slight. Results for Wave IV indicated, however, that the variables referring to the right to pay privately for better health care and education do correlate very highly before the analysis (r=0.775; p=0.000). Class clustering is able to moderate this correlation, but the bivariate residual for this pair in the model reported in Table 4 is still 1681.840.

To see how these correlations affect the results, we reexamine the model we previously estimated. We restrict the four highest bivariate residuals (the residual

for the *private healthcare just* and *private education just* pair and three others) to 0 and re-estimate the model. The resulting specification relaxes the assumption of conditional independence, possibly changing the relationship between the latent variable and indicators, and between covariates and indicators. Table 5 presents the results for our modified latent class analysis of Wave IV.

Table 5 Modified LCA of attitudes towards social policy in thirty-one countries (2009)

Model for Indicators

	Wald	p-value	R^2
government responsibility	2447.846	0.000	0.511
unemployment	770.994	0.000	0.254
poor	272.998	0.000	0.095
income differences	2143.097	0.000	0.392
top taxes	1120.467	0.000	0.258
progressive taxation	1079.735	0.000	0.183
private healthcare just	418.947	0.000	0.247
private education just	145.986	0.000	0.222

Model for Clusters

Intercept	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Wald	p-value
	-2.207	-2.547	-5.376	10.130	114.620	0.000
N	16093	2312	2370	2365		
Covariates						
sex						
male	-0.069	-0.080	0.204	-0.055	79.287	0.000
female	0.069	0.080	-0.204	0.055		
age	0.000	0.011	-0.006	-0.005	65.249	0.000
education						
no formal qualification	0.043	-0.265	0.340	-0.119	237.110	0.000
lowest formal qualification	0.048	0.171	-0.636	0.417		
above lowest qualification	0.180	0.334	-0.678	0.164		
higher secondary completed	-0.078	-0.062	0.004	0.135		
above higher secondary level, other qualification	-0.104	-0.064	0.415	-0.247		
university degree completed	-0.090	-0.113	0.555	-0.351		
GDP growth	0.313	0.254	0.217	-0.783	40.373	0.000
unemployment	0.001	-0.030	0.042	-0.013	15.175	0.002

employment status						
unemployed	0.003	0.011	-0.044	0.030	1.970	0.580
employed	-0.003	-0.011	0.044	-0.030		
income	-0.026	-0.272	0.343	-0.046	194.407	0.000
relative redistribution	0.071	0.107	0.063	-0.240	238.113	0.000
GDP	0.000	0.000	0.000	0.000	154.670	0.000
partisanship						
far left	-0.037	0.746	-1.201	0.492	736.296	0.000
left	0.333	0.586	-0.556	-0.363		
center	0.073	-0.049	0.207	-0.231		
right	0.056	-0.508	1.214	-0.761		
far right	0.154	-0.096	0.751	-0.809		
other, not specified	-0.745	-0.666	-0.776	2.187		
no party preference	0.166	-0.012	0.361	-0.516		
church attendance						
several times per week	-0.255	-0.068	-0.209	0.533	120.067	0.000
Once a week	0.145	-0.010	0.285	-0.421		
2 or 3 times a month	0.115	-0.069	0.303	-0.349		
Once a month	0.052	-0.199	0.017	0.130		
Several times a year	-0.016	0.143	-0.227	0.100		
Once a year	-0.004	-0.002	-0.118	0.124		
less than once a year	0.079	-0.037	0.103	-0.145		
never	-0.115	0.242	-0.154	0.028		
Overall N	23,426					

As Table 5 indicates, our software is now able to compute the Wald test statistic and its associated p-value for *GDP per capita*. Coefficients for *age*, *unemployment* and *unemployed* status have also turned positive, while the coefficient for *employed* is now negative. While *employed* and *unemployed* display the same signs as in the previous two waves, *employment status* overall has lost its statistical significance. Perhaps more importantly, R^2s for *private healthcare just* and *private education just* have significantly decreased, while *government responsibility* and *income differences* emerge once again as the indicators with the highest R^2s. This indicates that in its configuration, the latent class model for Wave IV is similar to the models for Waves I and II once the most glaring forms of conditional dependence have been properly handled.

With a refined model the number of observations by cluster can change, as Table 5 makes clear. Just as importantly, however, the BIC has declined in value. Having added one parameter (or restriction) to the model, researchers should check bivariate residuals again for additional parameters to restrict until all residuals

exhibit acceptable values. As more residuals are set to 0 and the ones left unrestricted decrease in value, we obtain diminishing increases in model fit (as judged by progressively lower BIC values), and more stability in parameters (the size of indicator and covariate coefficients and their signs). Due to space constraints, we do not report these checks here.[16]

We re-estimated our models with earnings instead of family income and obtained similar results except for Wave I, which exhibits a four-cluster profile somewhat different from the one we had originally obtained. Most likely, this is because data on earnings is not available for the Netherlands, and the country thus drops out of the estimation. We also experimented with a slightly modified form of latent class analysis, latent class factor analysis (Magidson & Vermunt 2001). LCFA is a form of exploratory factor analysis that conceives of attributes (self-interest and solidarism for example) as dichotomous latent factors rather than distinct classes. Instead of four latent classes, we would speak then of two dichotomous latent factors "with fixed and equidistant category scores" (Kankaraš et al. 2011, p. 284). While this allows individuals to have a position on each factor, LCFA achieves identification by omitting higher order interactions of the sort used previously. We thus found that our latent class models, judging by their lower BICs, fit the data better.

As a final check on our results, we estimated a model with data for all waves pooled into a single analysis. Since there are only four indicators common to all waves – *government responsibility, income differences, top taxes, and progressive taxation* – these are the only variables available to proxy for the latent construct. As before, we estimate models with 2, 3, and 4 latent classes. This time, we are able to work with thirty-two countries containing 32,678 observations.[17] Once again, a model with 4 classes fits the data best, judging by the BIC test statistic. As expected, the resulting class profile falls somewhere between the profiles for waves II and IV, with a class of individuals displaying a mixed set of attitudes.

We repeated the analyses with ISSP data from the Role of Government module. This questionnaire has the advantage of offering questions similar to the ones used here, in addition to questions on whether it should be the government's responsibility to provide decent housing for all and to care for seniors. Although we observed four classes underlying responses to these indicators, we were unable to obtain results similar to the ones just reported. The reason is most likely that the Role of Government module did not include questions asking people about their

16 Because bivariate residuals are smaller in the case of Waves I and II, we refrain from presenting refined versions of those models here.

17 Due to space constraints, we do not report details for this exercise here, but results are available upon request.

assessment of the status quo.[18] Because those questions prime answers to questions about absolute preferences, they cannot be separated empirically from indicators about abstract attitudes.

Putting all three waves together, we find that for two of the waves (Wave I and IV), the coefficient on income is negative, as one would expect, but for Wave II it is positive. We also see that signs for some macro-level variable coefficients are not stable across waves. Income redistribution and economic growth are sometimes associated with less (more) solidarism and unemployment with more (less) solidarism. These findings raise an important question: why are the effects of some variables inconsistent across waves? Rather than attempt to generalize when such generalizations are not warranted, we conclude that there is much about the relationship between personal/family income and macro-level variables that we still do not understand, particularly for developing and/or newer democracies such as those surveyed in Wave IV.

Dimick et al. (2017, p. 386) found evidence that "an increase in macro-inequality will lead to a larger increase in support for redistribution from the rich than from the poor". This occurs, they posit, because "an increase in redistribution aimed at reducing inequality is less costly (in welfare terms) to a richer person than to a poorer person" (i.e., the wealthy value an additional dollar of consumption less than the poor).[19] Haggard et al. (2013, p. 113) found, however, that in the developing world, "inequality has limited effects on demands for redistribution and may even dampen them." Others relate support for redistribution to its visibility (Gingrich 2014). Finally, some point out that spending on benefits locks some recipients into coalitions in favor of continued benefits (e.g. Timmons 2005).

Contradicting claims may reflect the reality that some variables, income in particular, are measured with error. Another possibility is that the effects of macro-level variables on attitudes differ between the more settled environments of developed countries and the more fluid situation we find in less developed ones. More generally, we believe that if people were fully informed, they would have no problems grasping the "inter-temporal trade-off between current and future income" that social policies entail (Barber et al. 2013, p. 1157). The cross-sectional nature of our research does not allow us to explore how stable over time the effects of these variables are, but it does allow us to realize that when care has been taken to specify the proper model, the relationship between covariates and the latent variable may not be the same across countries and/or waves.

---

18 There is no prompt in any of the surveys querying respondents about pro-poor policies specifically. These policies figure prominently in the welfare states of all advanced democratic nations.

19 See also Dimick et al. (2018).

Conclusion

This article has made a major contribution to the comparative political economy literature. As stated at the outset, solidarism, or care about the well-being of others, is usually treated as any attitude that cannot be explained using standard assumptions about self-interest. We have shown that this is not the case. Our empirical model provides strong support for the notion that solidarism is a coherent orientation among certain members of the public and it may or may not stem from "objective" indicators of wellbeing such as (relative) income, employment status, and education. Specifically, the model allowed us to measure these attitudes, thus helping overcome the by now stale divide between scholars who emphasize self-interested considerations over solidaristic behavior or vice-versa. We were able to do this while acknowledging the complexity of the relationship between contextual variables such as income redistribution and individual attitudes.

In addition to putting solidarism on a firmer empirical footing, this article made three other important contributions to the literature. First, we established some conceptual clarity regarding social policy preferences and how to measure them in a valid and reliable way. Second, we sorted through the thicket of how abstract preferences regarding the welfare state and information about its performance can affect each other. Finally, we showed how latent manifestations of these preferences might not be equivalent across countries.

We applied mixture modeling (LCA) to three waves of the International Social Survey Programme's module on social inequality. Our key findings are that preferences towards the market and the role of government in the economy form four distinct clusters of individuals that we refer to as "moderate altruists", "moderate egoists", "extreme altruists", and "extreme egoists". These clusters tend to be homogenous with respect to both abstract notions of the role of government in the economy as well as about evaluations of actual performance. We do find, however, one notable exception in the last two survey waves, as one class consists of individuals who are solidaristic with respect to some indicators, but self-interested with respect to others.

Looking at differences in results between waves, it appears as if attitudinal classes are context specific. We would expect the particular countries and indicator variables we study to affect class configurations. Our pooled analysis revealed, however, a configuration of classes across waves that is similar to the configurations found within them despite the smaller number of indicators used and heterogeneity introduced by pooling countries. There is something to be gained then from seeing latent classes as capturing four distinct types of attitudes that are fundamentally similar across units of analysis.

In future work, scholars should provide better accounts of why certain variables differ in their effects on attitudes across countries. The literature abounds with

claims about the relationship between variables such as inequality and redistributive preferences, but these works usually presume that effects are uniform across units while leaving direct country effects unexamined. As we have shown, even with an appropriate specification, such assumptions leave much to be explained. It is our hope that in the future, scholars not only measure attitudes more accurately, but also explain them better.

Literature

Alemán, J., & Woods, D. (2016). Value orientations from the World Values Survey: How comparable are they cross-nationally? *Comparative Political Studies* 49(8), 1039-1067.

Alesina, A., & Angeletos, G. (2005). Fairness and redistribution: US vs. Europe. *American Economic Review,* 95(3), 913-935.

Alesina, A., & Giuliano, P. (2011). Preferences for redistribution. In J. Benhabib, A. Bisin & M. O. Jackson (Eds.), *Handbook of social economics* (pp. 93-132). North Holland: Elsevier.

Alt, J., & Iversen, T. (2017). Inequality, labor market segmentation, and preferences for redistribution. *American Journal of Political Science,* 61(1), 21-36.

Arts, W., & Gelissen, J. (2001). Welfare states, solidarity and justice principles: Does the type really matter? *Acta Sociologica,* 44(4), 283-299.

Barber IV, B., Pablo, B., & Wibbels, E. (2013). The behavioral foundations of social politics: Evidence from surveys and a laboratory democracy. *Comparative Political Studies,* 46(10), 1115-1189.

Berlin, K. S., Williams, N. A., & Parra, G. R. (2014). An introduction to latent variable mixture modeling (part 1): Overview and cross-sectional latent class and Latent Profile analyses. *Journal of Pediatric Psychology,* 39(2), 174-187.

Bosancianu, C. M. (2017). A growing rift in values? Income and educational inequality and their impact on mass attitude polarization. *Social Science Quarterly,* 98(5), 1587-1602.

Breznau, N. (2010). Economic equality and social welfare: Policy preferences in five nations. *International Journal of Public Opinion Research,* 22(4), 458-484.

Breznau, N. & Hommerich, C. (2019). No generalizable effect of income inequality on public support for governmental redistribution among rich democracies 1987-2010, *Social Science Research*. doi: https://doi.org/10.1016/j.ssresearch.2019.03.013.

Cavaillé, C., & Trump, K. (2015). The two facets of social policy preferences. *The Journal of Politics,* 77(1), 146-160.

Choi, G. (2019). Revisiting the redistribution hypothesis with perceived inequality and redistributive preferences, *European Journal of Political Economy.* doi: https://doi.org/10.1016/ j.ejpoleco.2018.12.004.

De La O, A. L., & Rodden, J. A. (2008). Does religion distract the poor? Income and issue voting around the world. *Comparative Political Studies,* 41(4-5), 437-476.

Dimick, M., Rueda, D., & Stegmueller, D. (2017). The altruistic rich? Inequality and other-regarding preferences for redistribution. *Quarterly Journal of Political Science,* 11(4), 385-439.

Dimick, M., Rueda, D., & Stegmueller, D. (2018). Models of other-regarding preferences, inequality, and redistribution. *Annual Review of Political Science,* 21, 441-460.

Dion, M., & Birchfield, V. L. (2010). Economic development, income inequality, and preferences for redistribution. *International Studies Quarterly,* 54(2), 315-334.

Feenstra, R. C., Inklaar, R., & Timmer, M. P. (2015). The next generation of the Penn World Table. *American Economic Review,* 105(10), 3150-3182.

Fong, C. (2001). Social preferences, self-interest, and the demand for redistribution. *Journal of Public Economics* 82(2), 225–246.

Gimpelson, V., & Treisman, D. (2018). Misperceiving inequality. *Economics and Politics,* 30(1), 27-54.

Gingrich, J. (2014). Visibility, values, and voters: The informational role of the welfare state. *The Journal of Politics,* 76(2), 2-565.

Haggard, S., Kaufman, R. R., & Long, J. D. (2013). Income, occupation, and preferences for redistribution in the developing world. *Studies in Comparative International Development,* 48(2), 113-140.

Henry, K. H., & Muthén, B. (2010). Multilevel latent class analysis: An application of adolescent smoking typologies with individual and contextual predictors. *Structural Equation Modeling*, 17(2), 193–215.

International Social Survey Programme. (1987). *Social inequality module codebook*. Retrieved June 24, 2017, from the ISSP website: https://www.gesis.org/issp/modules/issp-modules-by-topic/social-inequality/1987/

International Social Survey Programme. (1992). *Social inequality II codebook*. Retrieved December 1, 2016, from the ISSP website: https://www.gesis.org/issp/modules/issp-modules-by-topic/social-inequality/1992/

International Social Survey Programme. (1999). *Social inequality III codebook*. Retrieved July 15, 2015, from the ISSP website: https://www.gesis.org/issp/modules/issp-modules-by-topic/social-inequality/1999/

International Social Survey Programme. (2009). *Social inequality IV codebook*. Retrieved July 15, 2015, from the ISSP website: https://www.gesis.org/issp/modules/issp-modules-by-topic/social-inequality/1999/

Iversen, T., & Soskice, D. (2001). An asset theory of social policy preferences. *American Political Science Review,* 95(4), 875-893.

Jensen, C., & Petersen, M. B. (2016). The deservingness heuristic and the politics of health care. *American Journal of Political Science,* 61(1), 68-83.

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica,* 47(2), 263-291.

Kankaraš, M., & Moors, G. (2009). Measurement equivalence in solidarity attitudes in Europe: Insights from a multiple-group latent-class factor approach. *International Sociology*, 24(4), 557-579.

Kelly, N. J., & Enns, P. K. (2010). Inequality and the dynamics of public opinion: The self-reinforcing link between economic inequality and mass preferences. *American Journal of Political Science,* 54(4), 855-870.

Kim, H., Huh, S., Choi, S., & Lee, Y. (2018). Perceptions of inequality and attitudes towards redistribution in four East Asian welfare states. *International Journal of Social Welfare,* 27(1), 28-39.

Kulin, J., Eger, M. A., & Hjerm, M. (2016). Immigration or welfare? The progressive's dilemma revisited. *Socius: Sociological Research for a Dynamic World*. https://doi.org/10.1177/2378023116632223.

Kunovich, R. M. (2009). The sources and consequences of national identification. *American Sociological Review,* 74(4), 573-593.

Larsen, C. A. (2008). The institutional logic of welfare attitudes: How welfare regimes influence public support. *Comparative Political Studies,* 41(2), 145-168.

Magidson, J., & Vermunt, J. K. (2001). Latent class factor and cluster models, bi-plots, and related graphical displays. *Sociological Methodology*, 31, 223-264.

Masyn, K. E. (2013). Latent class analysis and finite mixture modeling. In T. D. Little (Ed.), *The Oxford handbook of quantitative methodology in psychology, vol. 2: Statistical analysis* (pp. 551-661). Oxford University Press: Oxford.

Meltzer, A. H., & Richard, S. F. (1981). A rational theory of the size of government. *The Journal of Political Economy,* 89(5), 914-927.

Midtbø, T. (2017). Democracy and the demand for government redistribution: A survey analysis. *European Journal of Political Research*. doi.org/10.1111/1475-6765.12253.

Moene, K. O., & Wallerstein, M. (2001). Inequality, social insurance, and redistribution. *American Political Science Review,* 95(4), 859-874.

Moene, K. O., & Wallerstein, M. (2003). Earnings inequality and welfare spending: A disaggregated analysis. *World Politics,* 55(4), 485-516.

Moors, G. (2004). Facts and artifacts in the comparison of attitudes among ethnic minorities: A multigroup latent class structure model with adjustment for response style behavior. *European Sociological Review*, 20(4), 303-320.

Nagerkelke, E., Oberski, D. L., & Vermunt, J. K. Goodness-of-fit of multilevel latent class models for categorical data. *Sociological Methodology*, 46(1), 252-82.

Niehues, J. (2014). *Subjective perceptions of inequality and redistributive preferences: An International comparison*. Unpublished manuscript.

Oberski, D. L. (2016). Mixture models: Latent profile and latent class analysis. In J. Robertson, & K. Maurits (Eds.), *Modern statistical methods for HCI* (pp. 275-287). Springer, Cham.

O'Grady, T. (2017). How do economic circumstances determine preferences? Evidence from long-run panel data. *British Journal of Political Science*. doi:10.1017/S0007123417000242.

Rehm, P. (2011). Social policy by popular demand. *World Politics,* 63(2), 271-299.

Rehm, P., Hacker, J. S., & Schlesinger, M. (2012). Insecure alliances: Risk, inequality, and support for the welfare state. *American Political Science Review,* 106(2), 386-406.

Roosma, F., Gelissen, J., & van Oorschot, W. (2013). The multidimensionality of welfare state attitudes: A European cross-national study. *Social Indicators Research*, 113(1), 235–255.

Roosma, F., van Oorschot, W., & Gelissen, J. (2014). The preferred role and perceived performance of the welfare state: European welfare attitudes from a multidimensional perspective. *Social Science Research*, 44(C), 200-210.

Rueda, D. (2007). *Social democracy inside out: Partisanship and labor market policy in industrialized democracies*. Oxford: Oxford University Press.

Scheve, K., & Stasavage, D. (2006). Religion and preferences for social insurance. *Quarterly Journal of Political Science,* 1(3), 255-286.

Schmidt-Catran, A. W. (2016). Economic inequality and public demand for redistribution: Combining cross-sectional and longitudinal evidence. *Socio-Economic Review,* 14(1), 119-140.

Solt, F. (2016). The standardized world income inequality database, version 5.1. *Social Science Quarterly,* 97(5), 1267-1281.

Soroka, S. N., & Wlezien, C. (2010). *Degrees of democracy: Politics, public opinion, and policy*. Cambridge: Cambridge University Press.

Timmons, J. F. (2005). The fiscal contract: States, taxes, and public services. *World Politics,* 57(4), 530-567.

Trump, K. (2018). Income inequality influences perceptions of legitimate income differences. *British Journal of Political Science,* 48(4): 929-952.

VanHeuvelen, T. (2017). Unequal views of inequality: Cross-national support for redistribution 1985–2011. *Social Science Research,* 64, 43-66.

Vermunt, J. K. (2003). Multilevel latent class models. *Sociological Methodology*, 33(1), 213-239.

Vermunt, J. K., & Magidson, J. (2005). *Technical guide for latent GOLD choice 4.0: Basic and advanced*. Belmont, Massachusetts: Statistical Innovations Inc. Retrieved January 24, 2019, from the Statistical Innovations website: https://www.statisticalinnovations.com/wp-content/uploads/LGCusersguide.pdf.

Vermunt, J. K., & Magidson, J. (2016). *Technical guide for latent GOLD choice 5.1: Basic, advanced, and syntax*. Belmont, Massachusetts: Statistical Innovations Inc. Retrieved January 24, 2019, from the Statistical Innovations website: https://www.statisticalinnovations.com/wp-content/uploads/LGtecnical.pdf.

von Hippel, P. T., & Scarpino, S. V., Holas, I. (2016). Robust estimation of inequality from binned incomes. *Sociological Methodology,* 46(1), 212-251.

Wehl, N. (2018). The (ir)relevance of unemployment for labour market policy attitudes and welfare state attitudes. *European Journal of Political Research*. doi.org/10.1111/1475-6765.12274.

World Bank. (2017). World Development Indicators database. Retrieved August 4, 2017, from the World Development Indicators website: http://databank.worldbank.org/data/views/variableSelection/selectvariables.aspx?source=world-development-indicators.

Appendix A. Countries used in the analysis

Wave I	Wave II	Wave IV
Australia	Australia	Argentina
Austria	Austria	Australia
Germany	Germany	Austria
Netherlands	Norway	Belgium
Switzerland	United States	Bulgaria
United Kingdom		Czech Republic
United States		Denmark
		Estonia
		Finland
		France
		Germany
		Iceland
		Italy
		Japan
		Korea (South)
		Latvia
		New Zealand
		Norway
		Philippines
		Poland
		Portugal
		Slovakia
		Slovenia
		South Africa
		Spain
		Sweden
		Switzerland
		Ukraine
		United Kingdom
		United States
		Venezuela

methods, data, analyses | Vol. 14(1), 2020, pp. 91-126 DOI: 10.12758/mda.2019.07

Micro- and Macro-level Determinants of Participation in Demonstrations: An Analysis of Cross-national Survey Data Harmonized Ex-post

Marta Kołczyńska
Institute of Political Studies of the Polish Academy of Sciences

Abstract

This paper investigates micro- and macro-level determinants of participation in demonstrations worldwide, focusing on the role of resources and grievances across different democratic contexts. The analysis relies on a data set stemming from the ex-post harmonization of five international survey projects covering 100 countries between 1989 and 2009: Americas Barometer, Asia Europe Survey, European Values Study, International Social Survey Programme, and the World Values Survey. Results provide mixed support for previous findings and point to new insights. First, I find that the positive association between education and participation in demonstrations is stronger in democratic countries than in non-democracies, but there is no evidence of similar variation in the case of income. Second, the effect of trust in parliament is U-shaped, and more pronounced in non-democracies compared to democracies. Overall the findings indicate that the role of resources as well as disaffection with the political system in explaining participation in demonstrations depends on the political context, thus emphasizing the importance of incorporating both levels of analysis in theoretical and empirical models. The paper concludes with a discussion of the opportunities and challenges associated with ex-post harmonization of survey data.

Keywords: political participation, education, trust in state institutions, democracy, survey data harmonization, cross-national research

Political participation figures prominently in social science research as an avenue for citizens to communicate their views and to voice protest or support for political leaders. Much of cross-national research on political participation has been carried out in wealthy democracies, and this has shaped our understanding of the subject. Substantially less is known about determinants of participation in non-democratic and economically less-developed countries, and especially how they compare to those observed in democracies. This paper addresses long-standing questions in research on political behavior that pertain to the role of resources and grievances across political contexts (Cichocka et al., 2017; Dalton, Van Sickle, & Weldon, 2010), thus contributing to debates on the micro- and macro-determinants of political participation.

To extend the coverage and increase the representation of less democratic and economically developed countries, I rely on ex-post harmonized survey data from the Survey Data Recycling project (SDR, Slomczynski et al., 2017). Ex-post harmonization refers to procedures applied to existing data sets that were not created with comparability in mind, to transform original data sets in a way that enables us to analyze them as a single data source (Wolf, Schneider, Behr, & Joye, 2016). In this paper I use a subset of the SDR v.1 data (Slomczynski, Jenkins et al., 2017) consisting of five cross-national survey projects: Americas Barometer, Asia Europe Survey, European Values Study, International Social Survey Programme, and the World Values Survey. Together the data cover 100 countries between 1989 and 2009.

Results support prior findings about the positive association between individual resources – education and income – and participation in demonstrations, and further show that the association between education and participation is substantially stronger in democratic than in non-democratic countries. The findings related to the role of political trust are more complex and point to new insights: I find that

Acknowledgments

I would like to thank Kazimierz M. Slomczynski, J. Craig Jenkins, Joshua K. Dubrow, Przemek Powałko, and Robin Krauze for their help at various stages of preparing the paper, as well as the editor and the anonymous reviewers whose comments have greatly improved this manuscript.

This work was supported by the Institute of Philosophy and Sociology of the Polish Academy of Sciences (grant number 2017/S/05), the 2017 Silverman Research Support Award from the Department of Sociology at The Ohio State University, and the Visiting Scholar Program at the Chemnitz University of Technology.

Replication materials are available at https://osf.io/xadtn/.

Direct correspondence to

Marta Kołczyńska, Institute of Political Studies of the Polish Academy of Sciences, ul. Polna 18/20, 00-625 Warszawa, Poland
E-mail: kolczynska.1@osu.edu

trust is associated with the probability of demonstrating in a U-shaped way, and this pattern is more pronounced in non-democratic countries than in democracies.

The paper begins by presenting theoretical considerations on the role of resources, grievances, and the political context for political participation. Next, I describe the analytic strategy, including the harmonized survey data, as well as the process of arriving at the final data set for analysis. Since ex-post survey data harmonization is not a standard procedure in the social sciences, I discuss the harmonization strategy and the harmonization process of the survey variables used in this paper and the associated advantages and risks in some detail. After a description of the models, I turn to the results, followed by a discussion of the theoretical and methodological implications of the study. Beyond describing analyses using data from a specific harmonization project, the issues discussed in this paper are more broadly applicable to analyzing survey data characterized by varying quality and methodology.

Determinants of Participation in Demonstrations

Like most social phenomena, political participation results from a combination of factors, both individual and contextual, and is best analyzed in a multilevel theoretical framework (Dalton et al., 2010). Theories explaining political participation generally focus either on factors that enable participation by facilitating or removing barriers, or on factors that motivate participation by spurring opposition. According to the civic voluntarism model, political participation is enabled by the presence of resources (Berinsky, 2002; Brady, Verba, & Schlozman, 1995; Verba, Schlozman, & Brady, 1995). These resources can be of different kinds, including economic, but also civic skills, and the general expectation is that individuals with greater resources will be more likely to participate. The second approach, applied most often to contentious political participation, conceptualizes participation as motivated by grievances, which push people to go out into the streets and demand change (van Stekelenburg & Klandermans, 2013; Wilkes, 2004). Grievances are most frequently related to economic hardship, both absolute and relative (Klandermans, van der Toorn, & van Stekelenburg, 2008), but can also result from personal or political dissatisfaction (Muller, Jukam, & Seligson, 1982).

The Role of Resources: Income and Education

When focusing on economic well-being, the two theoretical approaches lead to contradictory hypotheses. While in the resource approach income is expected to be positively associated with the probability of participation, the grievance approach predicts a negative association. Empirical studies tend to support the first claim,

and find that individuals with higher income are more likely to participate in politics, both in conventional and unconventional activities (Loose & Jae, 2011; Marsh & Kaase, 1979; Quaranta, 2015).

However, not all resources are economic. Research has repeatedly shown that the better educated are more likely to engage in behaviors commonly associated with active citizenry: membership in voluntary associations (Putnam, 2000), protest behavior (Dalton et al., 2010), contacting politicians (Aars & Strømsnes, 2007), and reporting crimes to the police (Botero, Ponce, & Shleifer, 2012). Viewing education as a resource, explanations typically emphasize the cognitive costs of participation that are easier to overcome for educated individuals, who have better knowledge of the political system, can evaluate the performance of state institutions more accurately, and - in the case of under-performance - are better equipped to take action (Ceci, 1991; Marks, 2013; Winship & Korenman, 1997).

Political Trust as Grievance or as Resource

The relationship between political trust and political participation depends on the type of participation (see Gabriel, 2017, for a review). Protest behavior is considered more likely among individuals with low political trust, who reject conventional or "conformist" modes of participation and are more likely to engage in unconventional, elite-challenging activities, or withdraw from participation altogether (Citrin, 1974). In this sense, low trust constitutes a grievance against the political system. On the other hand, some level of trust in state institutions seems to be necessary for a person to engage in any political activity whose success depends on state responsiveness, which makes trust a resource that enables action (Cichocka et al., 2017).

Empirical studies provide mixed evidence. Some studies in Europe find a negative effect of political trust on protest activities, such as participation in demonstrations, boycotts, and signing petitions (Braun & Hutter, 2016; Kaase, 1999; Marien & Christensen, 2013; Marien & Hooghe, 2011). Another analysis of data from European countries found a positive effect of trust in the national parliament on "soft protest" (Dubrow, Slomczynski, & Tomescu-Dubrow, 2008). Yet another study, this time with a global scope, identified no effect of trust in parliament on protest participation (Dalton et al., 2010). Meanwhile, analyses by Cichocka and colleagues (2017) found a negative quadratic association between trust in state institutions and engaging in normative collective action. According to them, individuals having the least trust do not believe in the responsiveness of the state to protest, while those with very high trust exhibit a level of support which leaves little to protest against.

The Role of the Political Context

The same forms of political participation may have a different meaning depending on the political context. In democratic countries, activities such as attending demonstrations or wearing badges are legal, legitimate and generally safe, and have largely become part of the normal repertoire of politics (Dubrow et al., 2008; Newton & Montero, 2007). In authoritarian regimes, the same activities may be illegal and have a high chance of being repressed.

Considerations of contextual factors that shape political participation focus on the role of political opportunities, with theoretical expectations depending on whether the emphasis is on the enabling or on the motivating role of the context. Some scholars argue that openness of the political system, approval of mass participation, and responsiveness to protesters' demands, will encourage more participation (Eckstein & Gurr, 1975; Tarrow, 2011). Others claim that closed political systems that discourage civic engagement will increase protest participation if institutionalized channels are not accessible (Kitschelt, 1986). The differences in the mechanisms leading to political participation in democratic and non-democratic countries may result in a different composition of participants with regard to resources and grievances, as discussed above.

Hypotheses

If regime openness is generally associated with increased participation, it can be expected that the effect is stronger for individuals with more resources, and this is so for two reasons. First, these individuals are better equipped to identify and navigate the opportunities created by the political system. Second, in the case of low political openness and the potential for state repression, those with more resources have more to lose. Consequently, I expect that *the positive association between individual resources – education and income - and participation in demonstrations is stronger in more democratic countries than in less democratic ones* (Hypothesis 1a and 1b for the effects of education and income, respectively).

The role of political trust is also expected to vary across political contexts, in part due to the likely different nature of political participation following Kerbo's (Kerbo, 1982) distinction between movements of crisis and movements of affluence. *In non-democratic countries*, where collective action is discouraged or prohibited, *I expect high levels of political trust to be associated with regime loyalty and low propensity to participate in demonstrations* (Hypothesis 2). In these countries the distrust and dissatisfaction of citizens may accumulate and erupt in the form of mass demonstrations despite the fear of state repression, resulting in higher levels of participation in demonstrations among individuals with low political trust. On the other hand, following Cichocka et al. (2017), I expect that *in democratic countries*

the probability of participation is higher among individuals with medium levels of political trust than those with the highest and lowest trust levels (Hypothesis 3).

Analytic Strategy

Opportunities and Challenges of Survey Data Harmonization

Most empirical research of social and political issues focuses on democratic countries, largely because of the limited availability of survey data necessary to measure values, attitudes and participation from countries outside of the WEIRD (Western, Educated, Industrialized, Rich, and Democratic, cf. Henrich, Heine, & Norenzayan, 2010) zone (Kołczyńska, 2014; Slomczynski & Tomescu-Dubrow, 2006). Even beyond Europe, single survey programs do not include sufficient countries to analyze social and political phenomena on a global scale. To address this problem, the Survey Data Recycling project (SDR, Slomczynski & Tomescu-Dubrow, 2018; Słomczyński et al., 2016) set out to develop tools for combining data from many cross-national survey projects that were not intended to be comparable via ex-post harmonization, and for using the resulting harmonized data in substantive analyses.

The primary advantage of ex-post survey data harmonization is the increased coverage of countries and time points in the harmonized data set compared to data sets of single survey projects. This creates new opportunities for comparative research by enabling comparisons between countries and regions that are not covered by the same survey project, as well as over time. Associated costs are related to the increased methodological variation in the harmonized data set, including in the formulation of survey questions, the properties of response scales, or the sample types. All these factors can affect sample distributions of respondents' answers, and are a potential risk to the validity of conclusions stemming from analyses of ex-post harmonized data.

The SDR project proposes to address this issue by recording methodological information about the original (source) surveys as separate variables in the harmonized data set. This strategy is similar to the one employed by Milanovic in the *All the Ginis* data set of income inequality measures, where dummy variables distinguish between Gini coefficients that can potentially be incomparable (Milanovic, 2014). The methodological adjustment variables are of two types: harmonization controls and quality controls (Slomczynski & Tomescu-Dubrow, 2018). Harmonization controls are created during the harmonization of source variables and accompany each target (harmonized) variable. They capture properties of survey items that would be lost in the process of recoding or rescaling source variables into target variables, such as the length of response scales or characteristics of question wording. Harmonization controls are item-specific, i.e., they are constructed

individually for each target variable on the basis of the relevant methodological literature, which helps to identify the important features of items that are worth preserving, and following a review of source items in existing surveys to understand the variation in their design.

Quality control variables address the inter-survey variation in the methodology of the survey process or the quality of the data. Quality controls are either constructed on the basis of the available survey documentation (codebooks, study descriptions, technical reports) and describe important elements of the survey lifecycle, such as type of sample, or are derived from data records in the source data files to flag irregularities, such as duplicated records. Both types of control variables can be used in two ways: for the selection of surveys that meet pre-defined criteria or directly in the substantive models designed to test the hypotheses of the relationship between the chosen measures.

To sum up, while ex-post harmonization of surveys generally includes steps as presented below (cf. Granda et al., 2010; Wolf et al., 2016), the process employed in the SDR project includes additional stages marked in *italics*:

(1) concept definition:

a defining the target concept(s) to be measured with the survey variables, guided by the research question(s) and theoretical framework;

b based on this definition, developing a preliminary coding scheme or choosing a coding scale for the harmonized (target) variable;

(2) data preparation:

a identifying survey projects that meet the requirements regarding the presence of questions corresponding to the concepts identified in step 1.a, the target population and representativeness, and potentially other factors, and gathering their data and documentation;

b *examining the methodological variation among the gathered survey projects with regard to the design of the survey items of interest and the overall survey process on the basis of the survey documentation;*

c *describing surveys in terms of their methodology (e.g., sample type) and constructing survey quality indicators (e.g., the presence or absence of quality assurance procedures, proportion of duplicated cases);*

d identifying the candidate source variables, that is relevant question items in the gathered source surveys that correspond to the target concept(s) defined in step 1.a;

e *examining the variation in the design of the selected survey items given the literature on survey methodology and the effects of item design on respondents' answers;*

f *identifying relevant dimensions of variation between the survey items (e.g., related to item wording, response options or scales, position in the questionnaire, filtering) to be captured by harmonization control variables;*

g adjusting the coding scheme or scale of the harmonized (target variable) based on the observed variation in the survey items;

(3) harmonization:

a transforming (recoding) source variables into target variable(s) using the coding scheme established in step 2.g;

b *constructing harmonization control variables to capture the properties of source variables that would be lost in the process of recoding (e.g., details of original question wording or original length or direction of response scales), identified in step 2.f;*

(4) checking the target variable for errors and documentation of the whole process.

Data

The Survey Data Recycling v.1 data set (Slomczynski et al., 2017) stems from ex-post harmonization following procedures described above of selected variables from 22 international survey projects. The following sections describe the steps and decisions a researcher needs to make to prepare a data subset for analysis. These decisions deal with (1) the selection of surveys and cases from the harmonized SDR data set, and (2) accounting for methodological variation, including differences in survey quality and the variation in item design, across surveys.

Data Selection

Availability of variables. Not every national survey in the SDR data contains items measuring all the necessary concepts, so the availability of appropriate variables is the first criterion in the selection of surveys for analysis. Further constraints can be imposed by the selection of certain methodological features of some of these variables, which I discuss below.

Out of the 1721 national surveys in the SDR Master file, of the surveys carried out in 1989 or later, the period I will analyze in this paper[1], 646 national surveys have all the individual-level variables I want to include in models: participation in demonstrations (as the dependent variable), education, income, and trust in parliament (as individual-level independent variables), and age and gender (as controls).

1 Prior to 1989 data coverage is strongly skewed towards Western democracies, with little variation among the covered countries.

Sub-national samples. Some cross-national survey projects provide data for samples that are representative for sub-national populations, e.g., for a given region of the country. For example, the International Social Survey Programme typically has separate samples for East and West Germany. In the SDR v.1 data, national surveys are defined at the lowest possible level giving preference to sub-national samples where available. These include: Bosnia-Herzegovina (separate samples for the Federation of Bosnia and Herzegovina and Republika Srpska), Belgium (Flanders and Wallonia), Germany (East and West Germany), the United Kingdom (Great Britain and Northern Ireland), and Israel (separate samples for the Jewish and Arab populations). Most frequently, both split-up samples are provided, so that the entire territory of the given country is covered. To use data from split samples in an analysis of individuals nested in countries, I calculate additional weights proportional to the split samples' shares in the country's population. Occasionally, however, only one of the split samples is available, for example Belgium-Flanders in ISSP/2004, or Great Britain (without Northern Ireland) in ISSP/2014 or WVS/5. These "orphaned" samples are dropped from the analysis for two reasons. First, because of the lack of comparable contextual data on the level of the sub-national units. Second, because including them would mean that, for example, respondents from Belgium-Flanders are sometimes considered part of Belgium and sometimes – part of Belgium-Flanders, depending on the survey project, which creates difficulties for modeling. After eliminating the "orphaned" samples and combining sub-national samples into whole-nation samples, I am left with 628 surveys.

Selection on the properties of survey questions: **Participation in demonstrations.** The formulation of items that aim to capture political participation varies across projects, but they generally have the following form: *Have you performed [action type] in the last [time period]*?, where the time period ranges from "12 months" or "1 year" through 2, 3, 4, 5, 8, and 10 years to questions without any time frame (SDHT, 2017, pp. 79–84). Logically, the probability of a positive answer depends, among other things, on the time length the questions ask about. For any individual, the probability of participating in a demonstration in the last 5 years is greater or equal to the probability of participating in a demonstration in the last 12 months. This is why, when harmonizing data from different surveys, information

about the time span mentioned in the question must be recorded, and either used for the selection of data for analysis or accounted for when modeling the data[2].

It is unclear how the probability of participating in demonstrations depends on the time span due to at least two complicating factors. First, opportunities to demonstrate are not uniformly distributed in time. While occasional massive demonstration waves attract a substantial proportion of the population, there are also quieter times with fewer and less prominent events. Second, using retrospective questions introduces recall effects including temporal displacement, i.e. telescoping: respondents tend to report events earlier or later than they actually happened (Gaskell, Wright, & O'Muircheartaigh, 2000; Janssen, Chessa, & Murre, 2006; Neter & Waksberg, 1964). Human memory errors, including telescoping, but also omissions and overreporting, are related to age and education, as well as to the length of the time period and the frequency and salience of events (Ayhan & Işiksal, 2004). It is also possible that accuracy in reporting participation by respondents varies across cultures (Bernard, Killworth, Kronenfeld, & Sailer, 1984).

Since explicit modeling of recall effects across countries and cultures, time, and survey mode, is outside of the scope of this work, I restrict the data to surveys where questions asked about participation in demonstrations are without a time frame. This formulation is the most frequent among surveys in the SDR data set (SDHT, 2017), which provides sufficient variation in key country-level independent and control variables (quality of democracy, economic development), as well as large (and global) country coverage. Survey questions asking about participation in demonstrations "ever" can be understood as capturing a respondent's opinion about the legitimacy and perceived efficacy of the given form of participation, instead of actual past behavior in a temporal sense. Perceived efficacy is considered one of the main explanations for collective action (cf., Klandermans & van Stekelenburg, 2013, for a review), so this interpretation of the "ever" items is compatible with my theoretical framework.

Apart from the number of years in the question, items on participating in demonstrations in the selected subset also differ with regard to one other feature identified as potentially influencing respondents' answers: whether the question about participation in demonstrations mentioned other forms of participation in addi-

2 The formulation in the original questions can also vary within the same project wave, but across countries. One example are questions V100-V103 in World Values Survey Wave 5, which ask about recent participation in four activities: petitions, boycotts, demonstrations, and "other". According to the Master Questionnaire (WVS, 2005), the question is about participation in the last five years, but an analysis of country questionnaires reveals that in Hong Kong the question asked about the last 12 months, in Zambia about the last year. In Jordan there seems to be no indication of the time frame, and the question is missing from the questionnaire (and the data) from China, Colombia, Egypt, Guatemala, Iran, Iraq, Italy, Spain, and New Zealand. With the exception of China, all the other eight surveys contained the variable on participation in demonstrations "ever".

tion to demonstrations. For example, the project Asia Europe Survey asked about attending a "a protest, march or demonstration" (Inoguchi, 2008, p. 17). Mentioning other forms of participation next to demonstrations in the same question could be expected to yield a higher share of positive answers compared to a similar question that asks only about participation in demonstrations (Kołczyńska & Slomczynski, 2018), so surveys where questions about participation in demonstrations have this property are flagged with a control variable.

Non-unique records. Duplicate cases, or non-unique records, are a potential threat to data quality. In the SDR v.1 data set, the problem of duplicates was identified and analyzed by Slomczynski, Powałko, and Krauze (2017). Given the typical survey sample sizes and the number and types of survey items, encountering identical records can be considered a miracle or an error. Either way, they should be treated with suspicion.

In the SDR v.1 data set non-unique records are marked with a flag. Since non-unique records occur in the subset selected for analysis in this paper, I opted for the following strategy: surveys with more than five percent of duplicates are removed from the analysis, while in surveys with less than five percent of non-unique records, I drop all superfluous records following the recommendation of Sarracino and Mikucka (2017). The remaining subset consists of 332 national surveys.

Survey multiplets. Another issue that requires consideration are situations where more than one survey containing the necessary questions (after selecting the desired formulations) was carried out in the same country in the same year. Including them together in the models would increase the inequality in country coverage, and more frequently surveyed countries would weigh disproportionately on model estimates. To avoid this, from each country-year I selected only one sample with the largest proportion of cases without missing values on the variables of interest[3]. The resulting subset of the SDR data set used in the remainder of this paper includes data from 319 national surveys from five survey projects: editions 2004, 2006 and 2008 of the Americas Barometer (Americas Barometer, 2012), Asia Europe Survey (Inoguchi, 2001), editions 2, 3, and 4 of the European Values Study (European Values Study, 2011), International Social Survey Programme edition 2004 (ISSP Research Group, 2012), and editions 2, 3, 4, and 5 of the World Values Survey (World Values Survey, 2009). The list of countries by project edition is presented in Appendix A.

Accounting for Methodological Variation Across Surveys

As already mentioned, there is considerable methodological variation across survey projects, as well as between national surveys within the same project, with regard to many aspects of the survey process, as well as with regard to the resulting survey quality. The goal is to identify factors that can potentially affect the distribution of the variables of interest.

Item non-response. Item non-response, or the proportion of cases for which substantive responses for a given variable are not available, can be considered an indicator of the quality of the survey item (Groves, 1989), because it captures two aspects of item quality: the ability of the given item to elicit responses from respondents, and the extent to which the variable represents the variation in the measured characteristic in the population. To account for this, I include item non-response in the dependent variable as a control in the regression models.

Type of survey sample. All national surveys in SDR v.1 have samples coded on the basis of available documentation into seven categories: simple/stratified random sampling, multi-stage random sampling with individual register, multistage-random sampling with address register, samples with a random route component, samples with a quota component, and samples with inadequate or missing sampling descriptions. I include a control variable corresponding to the sample type to account for the possible systematic differences across national surveys relying on different types of samples.

Variables

Trust in parliament

The question about trust in the national parliament is the most popular survey items on political trust (Kołczyńska & Slomczynski, 2018). The harmonized variable "trust in parliament" used in this study was constructed in two steps (SDHT, 2017, pp. 49–55). First, variables originally coded on a descending scale were reversed so that in all variables lower scores mean less trust and higher scores - more trust. Second, variables were transformed into the target 0-10 scale. This transformation assumed that for scales shorter than 11 points each source value was assigned the mean of the corresponding range of values on the target 0-10 scale. For example, if the original scale had five points, the lowest value corresponds to the range between 0 and 2 on the 0-10 scale and was assigned the value of 1.

A control variable records the length of the original response scales in trust in parliament items in the source data, which in the case of the current analysis included questions with a 4-, 5-, and 7-point scales. Since the length of the original response scales influences the distribution of respondents' answers, and especially the differences between odd- and even-numbered scales can have an effect on the comparability of responses to the trust item, this control variable is included in models.

Education

To measure education, I use the target variable "Education level" from the SDR data set, which is harmonized on the basis of source variables indicating respon-

dents' educational attainment in terms of levels (SDHT, 2017, pp. 26–31). I recoded the levels into years by assigning to every level of education the mean number of years of schooling as suggested by UNESCO (2013; see also Słomczyński et al., 2016, pp. 181–182).

The SDR data set contains a second measure of education – "Years of schooling" – harmonized independently from "Education level" on the basis of questions asking about the number of years of schooling completed by the respondent, or the respondent's age at completion of education (SDHT, 2017, pp. 32–36). In surveys, in which "Education level" is not available, I used "Years of schooling" instead. Such cases are flagged with a control variable.

I chose to rely on "Education level" as the primary source of information about respondent's education and use "Years of schooling" to fill in gaps, because "Years of schooling" was in many cases calculated from responses to questions asking about respondent's age of completion of (taken together with respondent's year of birth or age), which is sensitive to the effects of returning to school by adults and more prone to errors.

Household income

The SDR data set does not contain any measure of individual economic status, so this variable was harmonized independently, in order to distinguish between the effects of economic status and of education (Kołczyńska & Powałko, 2019). The substantial variation in how the survey question about household income is asked (net or gross income, weekly, monthly, or annual income) and especially in how the responses are recorded (exact values, categories, quantiles) makes it hardly possible to harmonize household income in terms of assigning each respondent a monetary value in some common metric. Instead, the harmonized income variable was constructed by normalizing the original scale to the 0-100 range. Thus, the target variable "household income" captures the relative position of the respondent within the given national sample. It needs to be emphasized that this target variable does not allow for mean comparisons across samples.

Democracy

When looking at the whole spectrum of political regimes from autocracies to institutionalized democracies, the level of democracy may be treated as a less precise but appropriate indicator of the openness and responsiveness of the regime as well as of the probability of repression (Davenport & Armstrong, 2004). To measure democracy, I use Freedom House "Freedom in the World" ratings for Political Rights and Civil Liberties (Freedom House, 2016). The advantage of this indicator is its wide use in quantitative social science research, which lends credibility and offers global coverage. The Freedom House codes Political Rights and Civil Liberties on a scale from 1 to 7, where 1 represents the most and 7 the least freedom. I

use a sum of these measures, reversed so that the resulting variable is an indicator of democracy, not of the lack of democracy. The final variable is coded from 0 to 12, where 0 corresponds to the lowest, and 12 to the highest level of freedoms and liberties[4].

Control variables

In order to avoid attributing the effect of economic conditions to democracy, I control for GDP *per capita* using data from the World Bank's World Development Indicators (WDI, 2017)[5]. I also control for age and gender, which are known to be associated with political participation. Descriptive statistics for all individual-, macro-level, and methodological variables in their original metrics are presented in Table 1.

Models

To estimate the effects of micro- and macro-level factors, and their interactions, on reported participation in demonstrations, I estimate a series of three-level binary regression models, building up from the base model (Model 1) which takes the following form for individual *i* in country-year *j* in country *k*:

$$logit\left(participation_{ijk}\right) = \gamma_{000} + \gamma_{100} education_{ijk} + \gamma_{200} income_{ijk} +$$

$$\gamma_{300} trust_{ijk} + \gamma_{010} democracy_{jk} + \gamma_{x00} controls + r_{0jk} + u_{00k}$$

where γ_{000} is the global intercept, γ_{100}, γ_{200}, and γ_{300} are the coefficients for individual-level education, income, and trust in parliament, respectively, γ_{010} is the coefficient for country-year-level democracy, and γ_{x00} represents all coefficients for control variables at different levels, including substantive and methodological controls. Finally, r_{0jk} and u_{00k} are the random intercept terms.

Subsequent models each add an element of complexity. Model 2 adds a squared term for trust in parliament to test for quadratic effects of trust on participating in demonstrations. Models 3-5 include single cross-level interactions between the level of democracy and education, income, and trust in parliament, respectively.

---

4 The Czech Republic and Slovakia prior to their split-up in 1993 are assigned ratings from Czechoslovakia. Serbia and Montenegro in 1996 and 2001, and Kosovo in 2008 are assigned ratings from Yugoslavia for the respective years.

5 In rare cases when the value of GDP *per capita* was not available for the given country-year, the value from the adjacent year is used. Data for Taiwan are not available in the World Bank, so instead values from the International Monetary Fund's EconStats service are used: http://www.econstats.com/weo/CTWN.htm

Model 6 includes all three interaction terms, and the final Model 7 adds the methodological controls. In short, the models are built as follows:

Model 1: Base model;

Model 2: Model 1 + trust in parliament squared;

Model 3: Model 1 + education * democracy;

Model 4: Model 1 + income * democracy;

Model 5: Model 4 + trust in parliament * democracy + trust in parliament squared * democracy;

Model 6: Model 1 + education * democracy + income * democracy + trust in parliament squared + trust in parliament * democracy + trust in parliament squared * democracy;

Model 7: Model 6 + harmonization and methodological control variables.

In all analyses data are weighted with individual case weights provided in the source data sets and harmonized by SDR (SDHT, 2017, pp. 15–17). They are combined with weights proportional to the populations of sub-national regions where split samples were merged into national samples. In the analyses, trust in parliament is group-mean centered around the mean of the country-year, to estimate the effects of the relative level of trust within the country-year. I also include the country-year mean that captures the variation between country-years (Enders & Tofighi, 2007). All continuous variables are standardized by subtracting the mean and dividing by two standard deviations to facilitate comparisons of the magnitude of the coefficients within the same model (Gelman, 2008). While the values of the coefficients cannot be compared across models because of differences in the scale factor in non-linear probability models, their directions and significance levels remain informative (Breen, Karlson, & Holm, 2018).

To estimate the models I used the `glmer` command in the `lme4` package in R (Bates, et al., 2015), the `ggeffects` package (Lüdecke, 2018) to create the plots, and the `stargazer` package (Hlavac, 2018) for the tables. Multiple other R packages were used in the analysis: `rio` (Chan, Chan, Leeper, & Becker, 2018) to import and export data sets, `tidyverse` (Wickham, 2017) to clean and transform the data, `janitor` (Firke, 2019) to clean up variable names, `fastDummies` to recode categorical variables into sets of dummies (Kaplan, 2019), `democracyData` (Marquez, 2018) and `WDI` (Arel-Bundock, 2019) to download democracy and economic indicators, and `countrycode` (Arel-Bundock, Enevoldsen, & Yetman, 2018) to switch between country names and codes.

Table 1 Descriptive statistics of all variables included in the analysis in their original metrics.

Variable name	Mean / Proportion*	Std. dev.	Min	Max
Individual-level variables (n = 356,874)				
Participation in demonstrations	0.193		0	1
Trust in parliament	4.508	2.282	0.71	9.29
Trust in parliament (group mean centered)	0.001	2.097	-6.829	7.146
Education, years	10.587	4.363	0	18
Age, years	42.457	16.456	14	96
Female	0.512		0	1
Household income	38.722	27.165	0	100
Country-year-level variables (n = 319)				
Freedom House, reversed	9.395	2.902	0	12
GDP per capita, 000	20.627	15.061	1.088	94.900
GDP per capita, ln	9.611	0.888	6.992	11.461
Trust in parliament, sample mean	4.506	0.879	2.144	8.179
Year			1989	2009
Methodological variables (n = 319)				
Non-response on demonstrations	0.054	0.053	0.000	0.350
Question on demonstrations extended	0.160		0	1
Education filled with schooling years	0.213		0	1
Trust in parliament scale length				
4 points	0.784		0	1
5 points	0.103		0	1
7 points	0.113		0	1
Sample type				
No information	0.110		0	1
Insufficient information	0.232		0	1
Quota	0.313		0	1
Random route	0.154		0	1
Multistage address	0.078		0	1
Multistage individual	0.078		0	1
Single-stage	0.034		0	1
Survey project				
Americas Barometer	0.113		0	1
Asia Europe Survey	0.047		0	1
European Values Study	0.317		0	1
International Social Survey Programme	0.103		0	1
World Values Survey	0.420		0	1

* Proportions in the case of binary variables.

Results

Estimates of the conditional three-level models explaining individual participation in demonstrations are presented in Table 2. Model 1 is the baseline model with individual- and country-year-level covariates and controls, and random intercepts for all covariates. According to the model estimates, individual education and household income on average have a positive effect on participating in demonstrations, which is in line with the resource approach to explaining political participation. The standardized effect of education is about five times stronger than that of income, pointing to the role of non-economic resources in shaping participation decisions. The association between participation in demonstrations and the country's quality of democracy is also positive, in line with the expected role of political openness for political participation. The average linear effect of trust in parliament is weakly negative and not statistically significantly different from zero at the customary 0.05 level.

Coefficients for the individual-level control variables also largely conform to prior findings: participation in demonstrations is higher among men and the association with age forms an inverse-U, where the predicted probability of participating increases with age, peaks around 50 years, and declines to its minimum levels in old age. After controlling for the quality of democracy, economic development (GDP *per capita*) is negatively associated with the probability of demonstrating, while the effect of mean trust in parliament is positive suggesting that countries where individuals on average have more trust in the parliament see higher levels of participation in demonstrations.

Model 2 includes the quadratic term of trust in parliament. The coefficient is positive and statistically significant at the conventional level. The predicted association between trust in parliament and participation in demonstrations is hence U-shaped, where individuals with the lowest and highest levels of trust in parliament have the highest probability of participating in demonstrations, while individuals with medium levels – the lowest probability. This is the opposite pattern to the inverted-U that Cichocka et al. (2017) have found with the World Values Survey with a different operationalization of participation that took into account more activities.

Models 3, 4, and 5 add cross-level interactions of education, income, and trust in parliament, respectively, with the level of democracy. The significance of the interaction term in non-linear probability models is not a proper test of the interaction effect in terms of predicted probabilities (Mize, 2019), so the interactions are explored graphically below.

Model 6 includes all cross-level interactions – between individual education, income, and trust in parliament, and the country's level of democracy. The patterns

of associations remain stable with regard to their directions and magnitudes compared to Models 3, 4, and 5 with single interactions.

The final Model 7 adds methodological control variables of two types. The first are harmonization controls, which deal with variation in the design of original survey items: (a) an indicator for surveys where the original question about demonstrations also asked about another form of participation apart from demonstrations ("Question on demonstrations extended"), (b) information about the length of the original response scale in the "trust in parliament" items, and (c) a flag indicating whether the education variable substitutes schooling years for education levels. The second type includes other methodological controls: (a) the share of item non-response in the item about participation in demonstrations, and (b) the sample type employed in the given survey. While the coefficients for some of these controls are substantial, they only minimally change the effects of the individual-level covariates or the cross-level interactions. At the same time coefficients of macro variables – the level of democracy, GDP *per capita*, and mean trust in parliament – are affected much more, even if for the first two variables the directions and significance levels of the coefficients remain unchanged. The effect of mean trust in parliament becomes not statistically significant after adding control variables related to the length of the original response scales, which changes the substantive interpretation of the results. These changes in coefficients for macro-level predictors are not surprising given that harmonization controls and the sample type are measured on the level of the national survey corresponding to the country-year. As a result, including harmonization and quality controls will not likely change coefficients for individual-level predictors, especially if they are group-mean centered, but might affect coefficients for macro-level predictors in ways that may be difficult to interpret in substantive terms.

Table 2 Three-level logistic regression of individual-level participation in demonstrations.

Participation in demonstrations	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7
Education	0.846***	0.847***	0.927***	0.842***	0.845***	0.911***	0.914***
	(0.012)	(0.012)	(0.032)	(0.012)	(0.012)	(0.031)	(0.031)
Education * Democracy			0.421***			0.429***	0.431***
			(0.064)			(0.063)	(0.063)
Income	0.147***	0.147***	0.135***	0.170***	0.148***	0.164***	0.164***
	(0.009)	(0.009)	(0.010)	(0.020)	(0.010)	(0.018)	(0.018)
Income * Democracy				0.065		-0.016	-0.016
				(0.042)		(0.037)	(0.037)
Trust in parliament	-0.014	-0.016*	-0.022**	-0.016*	-0.035*	-0.041**	-0.041**
	(0.009)	(0.009)	(0.009)	(0.009)	(0.020)	(0.020)	(0.020)
Trust in parliament, squared		0.077***			0.069***	0.077***	0.076***
		(0.015)			(0.022)	(0.022)	(0.022)
Trust in parliament * Democracy					0.064	0.062	0.061
					(0.042)	(0.041)	(0.041)
Trust in parliament, squared * Democracy					-0.107**	-0.085*	-0.084*
					(0.045)	(0.045)	(0.045)
Democracy	0.626***	0.626***	0.605***	0.641***	0.659***	0.645***	0.518***
	(0.120)	(0.119)	(0.123)	(0.119)	(0.120)	(0.123)	(0.118)
Individual-level control variables							
Age	0.198***	0.197***	0.198***	0.206***	0.199***	0.204***	0.204***
	(0.011)	(0.011)	(0.011)	(0.011)	(0.011)	(0.011)	(0.011)
Age, squared	-0.455***	-0.456***	-0.433***	-0.456***	-0.456***	-0.445***	-0.444***
	(0.018)	(0.018)	(0.019)	(0.019)	(0.018)	(0.019)	(0.019)
Female	-0.351***	-0.350***	-0.353***	-0.352***	-0.351***	-0.353***	-0.353***
	(0.009)	(0.009)	(0.009)	(0.009)	(0.009)	(0.009)	(0.009)

Participation in demonstrations	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7
GDP per capita, ln	-0.308**	-0.304**	-0.333**	-0.320**	-0.309**	-0.340**	-0.262*
	(0.133)	(0.131)	(0.136)	(0.133)	(0.133)	(0.137)	(0.139)
Trust in parliament, mean	0.151*	0.149*	0.142*	0.158*	0.160*	0.149*	0.037
	(0.082)	(0.082)	(0.085)	(0.082)	(0.082)	(0.085)	(0.083)
Harmonization control variables							
Education filled with schooling years							-0.035
							(0.114)
Trust in parliament scale length (ref. 4 points)							
5 points							0.427***
							(0.092)
7 points							0.466**
							(0.228)
Question on demonstrations extended							0.179
							(0.129)
Quality control variables							
Non-response on demonstrations							0.009
							(0.070)
Sample type (ref. No information)							
Insufficient information							0.041
							(0.109)
Quota							-0.026
							(0.096)
Random route							-0.047
							(0.119)
Multi-stage address							-0.257*
							(0.138)

Participation in demonstrations	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7
Multi-stage individual							-0.119 (0.135)
Single-stage							0.110 (0.186)
Year	-0.006 (0.055)	-0.007 (0.055)	-0.003 (0.057)	-0.002 (0.055)	-0.016 (0.055)	-0.014 (0.057)	-0.133* (0.076)
Constant	-1.421*** (0.068)	-1.440*** (0.067)	-1.461*** (0.069)	-1.429*** (0.068)	-1.448*** (0.068)	-1.494*** (0.070)	-1.588*** (0.112)
Variance components							
Survey intercept	0.188	0.188	0.201	0.188	0.185	0.196	0.152
Education			0.239			0.222	0.223
Income				0.088		0.060	0.060
Trust in parliament					0.091	0.085	0.085
Trust in parliament, squared					0.052	0.052	0.052
Country intercept	0.357	0.356	0.374	0.359	0.369	0.385	0.413
Fit statistics							
Log Likelihood	-154,773	-154,761	-154,152	-154,506	-154,450	-153,722	-153,697
Akaike Inf. Crit.	309,572	309,549	308,333	309,042	308,936	307,487	307,460
Bayesian Inf. Crit.	309,712	309,700	308,495	309,204	309,130	307,725	307,816

Coefficients and standard errors in parentheses. * $p<0.05$, ** $p<0.01$, *** $p<0.001$.
N individuals = 356,874, N surveys = 319, N countries = 100.
Data source: Survey Data Recycling v.1, Harmonized Income Database v.1, Freedom House, World Bank, International Monetary Fund.

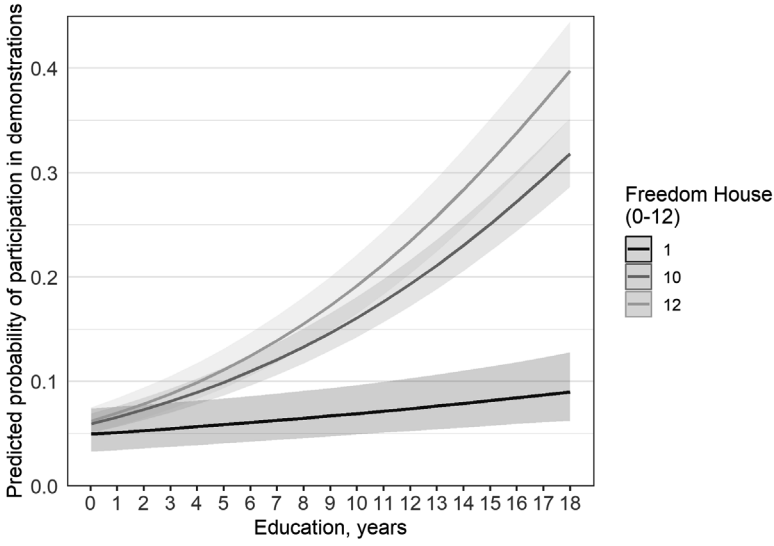


Figure 1 Predicted probability of participation in demonstrations by education and democracy (based on Model 7).

Predicted probabilities of participation in demonstrations illustrating the effects of individual education, income, and trust in parliament and their interactions with the level of democracy, with other covariates held at their means or at base levels for factors, are presented in Figures 1, 2, and 3. Figure 1 shows how the positive effects of education increase with increasing levels of democracy, in line with Hypothesis 1a. In the least democratic countries (Freedom House score equal to 1 on the scale from 0 to 12), the difference between the predicted probability of participation in demonstrations for those with no education and those with secondary education (12 years of schooling) is less than 3 percentage points, while in the most democratic countries (Freedom House score of 12) the difference is around 13 percentage points. Moving from secondary education (12 years) to tertiary education (16 years) corresponds to a change in predicted probability of demonstrating by 10 percentage points in democratic countries and by one percentage point in the least democratic countries.

Figure 2 presents the association between participation in demonstrations and household income at different levels of democracy, and shows that the effect of income on participating in demonstrations is positive at all levels of democracy, and is stronger the higher the more democratic the country. Moving from the lowest income to the highest income in non-democratic countries increases the probabil-

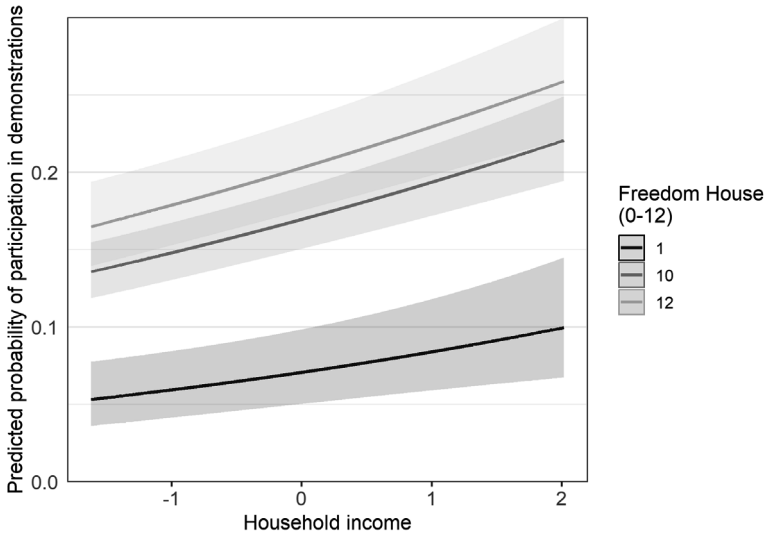


Figure 2 Predicted probability of participation in demonstrations by levels of income and democracy (based on Model 7).

ity of participation by less than 5 percentage points, while in the most democratic countries the corresponding change is by around 8 percentage points. These results need to be taken with a grain of salt given how imperfect the harmonized measure of household income is. It is possible that the observed interaction effect is due to differences in the measurement of income between surveys, or in the distribution of income between less and more democratic countries. Even if real, the difference in the magnitude of the effect of income by level of democracy is far smaller than of the effects of education, and the support for Hypothesis 1b is weak at best.

The predicted levels of participation in demonstrations depending on trust in parliament and by levels of democracy are presented in Figure 3, showing the U-shaped association between the probability of demonstrating in non-democratic countries. In these countries, the highest predicted probability of participating in demonstrations is for individuals with the lowest levels of trust in parliament at 0.13. Individuals with a medium-high level of trust in parliament have the lowest predicted probability of demonstrating of 0.065. The predicted probability increases for individuals with the highest level of trust in parliament to almost 0.1. In democratic countries the association is much flatter, and the difference between the lowest and the highest predicted probability of demonstrating is less than 2

percentage points. These results contradict the expectations stated in Hypotheses 2 and 3[6].

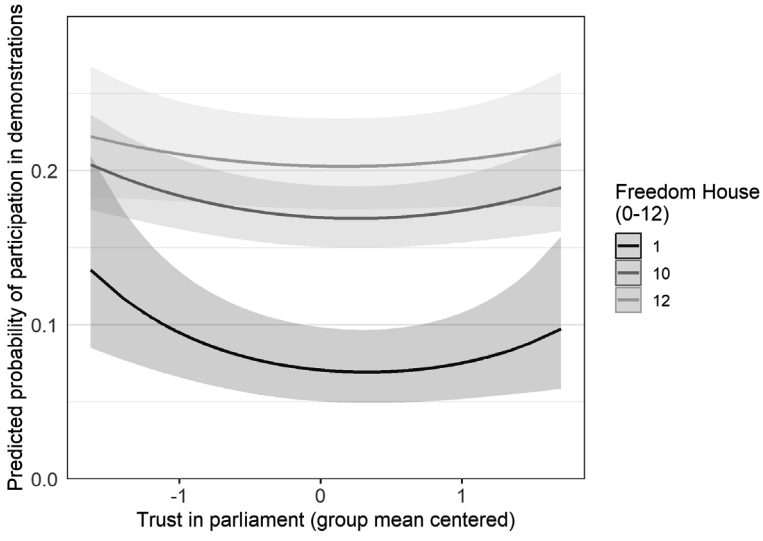


Figure 3 Predicted probability of participation in demonstrations by levels of political trust and democracy (based on Model 7).

Conclusion

In this paper I analyzed individual and contextual determinants of participation in demonstrations with data from 100 countries between 1989 and 2009, using ex-post harmonized data from five international survey projects. Results provide mixed support for previous findings and point to new insights. First, the analysis reveals systematic variation in the effects of education on participation in demonstrations: the effect of education on participation in demonstrations is positive and far stronger in democracies than in non-democracies. This might be because, while educated individuals are better at recognizing opportunities for meaningful participation and exploiting them, in non-democratic countries the awareness of limited chances for success might keep them from taking to the streets. Additionally, educated individuals who engage in protests in non-democratic countries face comparatively higher risks of state repression than in democracies. At the same time, while the association between income and participation in demonstrations is also positive, the magnitude of the effect and its variation across levels of democ-

racy are much weaker. These results confirm prior findings about the central role of education for political participation.

Further, I found that political trust is related to participation in demonstrations in a complex way: it is U-shaped, but the pattern is the strongest in the least democratic countries, and very weak in institutionalized democracies. If in non-democracies both the least and the most trusting citizens demonstrate the most, are they participating in the same demonstrations? Perhaps the demonstrations attended by individuals who are distrustful of the political regime indeed constitute protest, while in the case of individuals with high political trust in a non-democratic country, demonstrations could rather be in support of than against the state (cf. Hellmeier & Weidmann, 2019). Standard survey questions about participation in demonstrations do not distinguish between demonstrations for and demonstrations against the political system, and variation between countries might be exacerbated by linguistic differences in the meaning and connotations of the word "demonstration" or an alternative term used in the survey question. In general, verifying the validity of the assumption that participation in demonstrations, as measured in surveys, is a form of protest, could explain some of the mixed findings in the empirical literature on this topic.

The second goal of the paper is to provide an illustration of how survey data harmonized ex-post can be used in a substantive analysis. The approach to ex-post harmonization proposed in the SDR project consists in unifying the coding of original (source) variables that are identified as measuring the same concept by either mapping the original values onto a common coding scheme or by rescaling the responses to a common range, in addition to constructing auxiliary variables to record selected properties of the source variables. In this paper, I showed how the harmonized data created in the SDR approach can be applied to a concrete research problem.

Data from ex-post harmonization, such as performed in the SDR project, are not without limitations. First, while the SDR data set increases country coverage through harmonization of survey data from different cross-national surveys, the inequality in country coverage persists, and the time series for less developed countries remain short and sparse, especially after selecting a subset of the data set with the necessary harmonized variables. Second, the harmonization of variables requires that survey projects include the same or very similar questions. As a consequence, analyses are limited by the number of available harmonized variables enabling the estimation of fairly modest models. Such models can identify only broad patterns of associations for further examination with richer data sets.

Third, the process of harmonization as employed in the SDR project entails information loss and may introduce bias when response categories are collapsed, or when original responses measured with ordinal rating scales are treated as continuous and rescaled. Overall, ex-post harmonization introduces harmonization error

with unknown properties. The SDR framework aims to mitigate this by constructing control variables that are supposed to capture the aspects of question design that are lost in the process of standardization as well as the methodological and quality variation between the different surveys, but the extent to which this is successful is yet to be adequately examined. As this paper shows, analyses focusing on individual-level predictors or cross-level interactions yield stable results whether or not methodological and quality controls are applied. At the same time caution is advised when analyzing the effects of macro-level variables, as they can be correlated with the methodological and quality controls in spurious ways, which would have an effect on model coefficients, and could result in interpreting data artifacts in substantive terms.

The more general question is how to balance the costs and benefits of harmonizing survey data that were not *a priori* intended for joint analysis. On the one hand, research in cross-cultural survey methodology has led to the development of standards and guidelines that greatly improve the comparability of cross-national survey data and has demonstrated how disregarding these standards during the survey process may hurt the comparability of the resulting data (Survey Research Center, 2016). This research is focused on improving future data collection efforts, and implicitly questions the value of cross-national data sets constructed from surveys that were collected without careful ex-ante considerations of comparability at all stages of the survey process. On the other hand, surveys carried out over the last several decades in many countries worldwide are valuable as historical evidence, and researchers may be tempted to harmonize all of them together regardless of their known or suspected limitations.

In the middle ground between the extreme positions of dismissing any ex-post harmonization and combining all surveys regardless of their quality, there seem to be two main questions. The first pertains to the minimum standards for including a survey data set in a comparative analysis, with the discussion likely organized around issues related to the quality of the sample and of the measurement. The second question deals with methods of modeling survey data stemming from ex-post harmonization, and limitations to the types of statistical analyses that can be performed with such data. Efforts aimed at formulating recommendations in response to both questions would benefit from a comprehensive framework to evaluate survey quality. The quality assessment approach in the SDR project constructed quality indicators in three dimensions: quality of data, quality of documentation, and correspondence between the data and the documentation (Slomczynski & Tomescu-Dubrow, 2018). Others have attempted to assess the quality of survey samples on the basis of internal and external criteria of representativeness (Jabkowski & Cichocki, 2019; Kohler, 2007; Kołczyńska, Cichocki, & Jabkowski, 2019). While promising, these attempts face limitations related to the data and doc-

umentation, and further work is also needed in the area of measurement equivalence with survey data harmonized ex-post.

Finally, while debating the limitations of ex-post survey harmonization, it is worth remembering that many of the same challenges apply – although arguably to a lesser extent – to data within a single cross-national survey project, which goes largely unaddressed in empirical studies. Cross-national survey projects often collect data following different protocols in different countries, and these protocols change from edition to edition. Other aspects of survey quality, including documentation standards and survey outcome rates, also vary – within the same project – across countries and change over time (Jabkowski, 2018; Kołczyńska & Slomczynski, 2018; Oleksiyenko, Wysmułek, & Vangeli, 2018). Further, variables available in single cross-national survey data sets also face limitations with regard to their comparability and interpretability (Donnelly & Pop-Eleches, 2018), and survey projects themselves often ex-post harmonize the coding of socio-demographic variables, a process prone to errors.

The discussion about the consequences of combining survey data collected following different standards and procedures, and about minimum thresholds for data quality, is thus not limited to ex-post harmonized data from different projects, but also applies to analyses of data from single cross-national survey projects. Ultimately, it is up to the researcher to decide which surveys are of sufficiently high quality to be included in the analysis. Since most researchers are secondary data users, the availability and high information content of survey documentation is of utmost importance in this process.

References

Aars, J., & Strømsnes, K. (2007). Contacting as a Channel of Political Involvement: Collectively Motivated Individually Enacted. *West European Politics*, *30*(1), 93–120. https://doi.org/10.1080/01402380601019704

Americas Barometer. (2012). Americas Barometer Merged 2004-2012 Rev 1.5. Latin American Public Opinion Project. Retrieved from www.LapopSurveys.org

Arel-Bundock, V. (2019). WDI: World Development Indicators (World Bank). Retrieved from https://cran.r-project.org/package=WDI

Arel-Bundock, V., Enevoldsen, N., & Yetman, C. J. (2018). countrycode: An R package to convert country names and country codes. *Journal of Open Source Software*, *3*(28), 848. Retrieved from https://doi.org/10.21105/joss.00848

Ayhan, H. Ö., & Işiksal, S. (2004). Memory recall errors in retrospective surveys: A reverse record check study. *Quality and Quantity*, *38*(5), 475–493. https://doi.org/10.1007/s11135-005-2643-7

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Berinsky, A. J. (2002). Silent Voices: Social Welfare Policy Opinions and Political Equality in America. *American Journal of Political Science*, *46*(2), 276.

Bernard, H. R., Killworth, P., Kronenfeld, D., & Sailer, L. (1984). The Problem of Informant Accuracy: The Validity of Retrospective Data. *Annual Review of Anthropology*, *13*(1), 495–517. https://doi.org/10.1146/annurev.an.13.100184.002431

Botero, J., Ponce, A., & Shleifer, A. (2012). Education and the Quality of Government. *NBER Working Paper Series*, (18119). Retrieved from http://www.nber.org/papers/w18119

Brady, H. E., Verba, S., & Schlozman, K. L. (1995). Beyond SES: A Resource Model of Political Participation. *American Political Science Review*, *89*(2), 271–294. https://doi.org/10.2307/2082425

Braun, D., & Hutter, S. (2016). Political trust, extra-representational participation and the openness of political systems. *International Political Science Review*, *37*(2), 151–165. https://doi.org/10.1177/0192512114559108

Breen, R., Karlson, K. B., & Holm, A. (2018). Interpreting and Understanding Logits, Probits, and Other NonLinear Probability Models. *Annual Review of Sociology*, *44*(1), 39–54. https://doi.org/10.1146/annurev-soc-073117-041429

Ceci, S. J. (1991). How Much Does Schooling Influence General Intelligence and Its Cognitive Components? A Reassessment of the Evidence. *Developmental Psychology*, *27*(5), 703–722.

Chan, C., Chan, G. C. H., Leeper, T. J., & Becker, J. (2018). rio: A Swiss-army knife for data file I/O. Retrieved from https://cloud.r-project.org/web/packages/rio/index.html

Cichocka, A., Górska, P., Jost, J. T., Sutton, R. M., & Bilewicz, M. (2017). What Inverted U Can Do for Your Country: A Curvilinear Relationship Between Confidence in the Social System and Political Engagement. *Journal of Personality and Social Psychology*. https://doi.org/10.1037/pspp0000168

Citrin, J. (1974). Comment: The Political Relevance of Trust in Government. *American Political Science Review*, *68*(3), 973–988. https://doi.org/10.2307/1959141

Dalton, R. J., Van Sickle, A., & Weldon, S. (2010). The individual-institutional nexus of protest behaviour. *British Journal of Political Science*, *40*(1), 51–73. https://doi.org/10.1017/S000712340999038X

Davenport, C., & Armstrong, D. A. (2004). Democracy and the Violation of Human Rights: A Statistical Analysis from 1976 to 1996. *American Journal of Political Science*, *48*(3), 538–554. https://doi.org/10.1111/j.0092-5853.2004.00086.x

Donnelly, M. J., & Pop-Eleches, G. (2018). Income Measures in Cross-National Surveys: Problems and Solutions. *Political Science Research and Methods*, *6*(2), 355–363. https://doi.org/10.1017/psrm.2016.40

Dubrow, J. K., Slomczynski, K. M., & Tomescu-Dubrow, I. (2008). Effects of Democracy and Inequality on Soft Political Protest in Europe: Exploring the European Social Survey Data. *International Journal of Sociology*, *38*(3), 36–51. https://doi.org/10.2753/IJS0020-7659380302

Eckstein, H., & Gurr, T. R. (1975). *Patterns of Authority: A Structural Basis for Political Inquiry*. New York: Wiley.

Enders, C. K., & Tofighi, D. (2007). Centering Predictor Variables in Cross-Sectional Multilevel Models: A New Look at an Old Issue. *Psychological Methods*, *12*(2), 121–138. https://doi.org/10.1037/1082-989X.12.2.121

European Values Study. (2011). European Values Study Longitudinal Data File 1981-2008 (EVS 1981-2008). ZA4804 Data file Version 2.0.0. Cologne: GESIS Data Archive. https://doi.org/10.4232/1.11005

Firke, S. (2019). janitor: Simple Tools for Examining and Cleaning Dirty Data. Retrieved from https://cran.r-project.org/package=janitor

Freedom House. (2016). Freedom in the World, 1973-2016. Retrieved from https://freedomhouse.org/report-types/freedom-world

Gabriel, O. W. (2017). Participation and Political Trust. In S. Zmerli & T. van der Meer (Eds.), *Handbook on Political Trust* (pp. 228–241). Cheltenham and Northampton: Edward Elgar Publishing.

Gaskell, G. D., Wright, D. B., & O'Muircheartaigh, C. A. (2000). Telescoping of Landmark Events: Implications For Survey Research. *Public Opinion Quarterly*, *64*, 77–89.

Gelman, A. (2008). Scaling regression inputs by dividing by two standard deviations. *Statistics in Medicine*, *27*(15), 2865–2873. https://doi.org/10.1002/sim.3107

Granda, P., Wolf, C., & Hadorn, R. (2010, May 5). Harmonizing Survey Data. *Survey Methods in Multinational, Multiregional, and Multicultural Contexts*. https://doi.org/doi:10.1002/9780470609927.ch17

Groves, R. M. (1989). *Survey Errors and Survey Costs*. New York: Wiley.

Hellmeier, S., & Weidmann, N. B. (2019). Pulling the Strings? The Strategic Use of Pro-Government Mobilization in Authoritarian Regimes. *Comparative Political Studies*, On-line first. https://doi.org/10.1177/0010414019843559

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences*, *33*(2–3), 61–83. https://doi.org/10.1017/S0140525X0999152X

Hlavac, M. (2018). stargazer: Well-Formatted Regression and Summary Statistics Tables. Bratislava, Slovakia. Retrieved from https://cran.r-project.org/package=stargazer

Inoguchi, T. (2001). Asia Europe Survey (ASES): A Multinational Comparative Study in 18 Countries. Ann Arbor, MI: Inter-university Consortium for Political and Social Research. https://doi.org/10.3886/ICPSR22324.v1

Inoguchi, T. (2008). Asia Europe Survey (ASES): A Multinational Comparative Study in 18 Countries, 2001. Codebook. Inter-university Consortium of Political and Social Research. https://doi.org/10.3886/ICPSR22324.v1

ISSP Research Group. (2012). International Social Survey Programme: Citizenship - ISSP 2004. GESIS Data Archive, Cologne. ZA3950 Data file Version 1.3.0. https://doi.org/10.4232/1.11372

Jabkowski, P. (2018). *Surveys Quality Assessment Database (SQAD)*. Retrieved from https://www.researchgate.net/project/Surveys-Quality-Assessment-Database-SQAD

Jabkowski, P., & Cichocki, P. (2019). Within-household selection of target-respondents impairs demographic representativeness of probabilistic samples: evidence from seven rounds of the European Social Survey. *Survey Research Methods*, *13*(2), 167–180. https://doi.org/10.18148/srm/2019.v13i2.7383

Janssen, S. M. J., Chessa, A. G., & Murre, J. M. J. (2006). Memory for time: How people date events. *Memory & Cognition*, *34*(1), 138–147. https://doi.org/10.3758/BF03193393

Kaase, M. (1999). Interpersonal trust, political trust and non-institutionalised political participation in Western Europe. *West European Politics*, *22*(3), 1–21. https://doi.org/10.1080/01402389908425313

Kaplan, J. (2019). fastDummies: Fast Creation of Dummy (Binary) Columns and Rows from Categorical Variables. Retrieved from https://cran.r-project.org/package=fastDummies

Kerbo, H. R. (1982). Movements of "Crisis" and Movements of "Affluence." *Journal of Conflict Resolution*, *26*(4), 645–663. https://doi.org/10.1177/0022002782026004004

Kitschelt, H. P. (1986). Political Opportunity Structures and Political Protest: Anti-Nuclear Movements in Four Democracies. *British Journal of Political Science*, *16*(1), 57–85. https://doi.org/10.1017/S000712340000380X

Klandermans, B., van der Toorn, J., & van Stekelenburg, J. (2008). Embeddedness and Identity: How Immigrants Turn Grievances into Action. *American Sociological Review*, *73*(6), 992–1012. https://doi.org/10.1177/000312240807300606

Klandermans, B., & van Stekelenburg, J. (2013). *Social Movements and the Dynamics of Collective Action*. (L. Huddy, D. O. Sears, & J. S. Levy, Eds.), *The Oxford Handbook of Political Psychology*. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199760107.013.0024

Kohler, U. (2007). Surveys from inside: An assessment of unit nonresponse bias with internal criteria. *Survey Research Methods*, *1*(2), 55–67. https://doi.org/10.18148/srm/2007.v1i2.75

Kołczyńska, M. (2014). Representation of Southeast European Countries in International Survey Projects: Assessing Data Quality. *Ask: Research and Methods*, *23*(1), 57–78.

Kołczyńska, M., Cichocki, P., & Jabkowski, P. (2019). Quality evaluation of survey samples on the basis of external and internal criteria of representativeness. Evidence based on 1721 surveys from major cross-country projects. *8th Conference of the European Survey Research Association, Zagreb, 19 July 2019.*

Kołczyńska, M., & Powałko, P. (2019). Harmonized Income Dataset. Harvard Dataverse, V1, UNF:6:nFDl873ADY9y1j8Sia3bVA== [fileUNF]. Retrieved from https://doi.org/10.7910/DVN/UE7XIJ

Kołczyńska, M., & Slomczynski, K. M. (2018). Item Metadata as Controls for Ex Post Harmonization of International Survey Projects. In T. P. Johnson, B.-E. Pennell, I. A. L. Stoop, & B. Dorer (Eds.), *Advances in Comparative Survey Methods: Multinational, Multiregional, and Multicultural Contexts (3MC)* (pp. 1011–1034). Wiley.

Loose, K., & Jae, D. H.-S. (2011). Explaining Unequal Participation: The Differential Effects of Winter Weather on Voter Turnout. *Massachusetts Institute of Technology Political Science Department. Working Paper*, (No. 2011-13). Retrieved from https://ssrn.com/abstract=1802736

Lüdecke, D. (2018). ggeffects: Tidy Data Frames of Marginal Effects from Regression Models. *Journal of Open Source Software*, *3*(26), 772. https://doi.org/10.21105/joss.00772

Marien, S., & Christensen, H. S. (2013). Trust and Openness: Prerequisites for Democratic Engagement? In K. N. Demetriou (Ed.), *Democracy in Transition* (pp. 109–134). Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-30068-4_7

Marien, S., & Hooghe, M. (2011). Does political trust matter? An empirical investigation into the relation between political trust and support for law compliance. *European Journal of Political Research*, *50*(2), 267–291. https://doi.org/10.1111/j.1475-6765.2010.01930.x

Marks, G. N. (2013). *Education, Social Background and Cognitive Ability: The Decline of the Social*. New York: Routledge.

Marquez, X. (2018). democracyData: Access and manipulate most standard scholarly measures of democracy. Retrieved from https://github.com/xmarquez/democracyData

Marsh, A., & Kaase, M. (1979). Background of political action. In S. H. Barnes & M. Kaase (Eds.), *Political action. Mass participation in five western democracies* (pp. 97–136). Beverly Hills: Sage.

Milanovic, B. (2014). Description of All The Ginis Dataset. *World Bank, Research Department*. Retrieved from http://pubdocs.worldbank.org/en/728601472744842249/Description-of-AllGinis-Oct2014.pdf

Mize, T. (2019). Best Practices for Estimating, Interpreting, and Presenting Nonlinear Interaction Effects. *Sociological Science*, *6*, 81–117. https://doi.org/10.15195/v6.a4

Muller, E. N., Jukam, T. O., & Seligson, M. A. (1982). Diffuse Political Support and Antisystem Political Behavior: A Comparative Analysis. *American Journal of Political Science*, *26*(2), 240–264. https://doi.org/10.2307/2111038

Neter, J., & Waksberg, J. (1964). A Study of Response Errors in Expenditures Data from Household Interviews. *Journal of the American Statistical Association*, *59*(305), 18–55. https://doi.org/10.1080/01621459.1964.10480699

Newton, K., & Montero, J. R. (2007). Patterns of Political and Social Participation in Europe. In R. Jowell, C. Roberts, R. Fitzgerald, & G. Eva (Eds.), *Measuring Attitudes Cross-Nationally* (pp. 204–237). London: SAGE Publications, Ltd. https://doi.org/10.4135/9781849209458.n10

Oleksiyenko, O., Wysmułek, I., & Vangeli, A. (2018). Identification of Processing Errors in Cross-national Surveys. In Timothy P. Johnson, B.-E. Pennell, I. A. L. Stoop, & B. Dorer (Eds.), *Advances in Comparative Survey Methods: Multinational, Multiregional, and Multicultural Contexts (3MC)* (pp. 985–1010). Wiley.

Putnam, R. (2000). *Bowling Alone: The Collapse and Revival of American Community.* New York: Simon & Schuster.

Quaranta, M. (2015). *Political Protest in Western Europe*. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-22162-5

Sarracino, F., & Mikucka, M. (2017). Bias and effiency loss in regression estimates due to duplicated observations: a Monte Carlo simulation. *Survey Research Methods*, *11*(1), 17–44. https://doi.org/10.18148/srm/2017.v11i1.7149

SDHT (Survey Data Harmonization Team). (2017). Master file documentation. SDR Master Box Version 1. https://doi.org/10.7910/DVN/VWGF5Q

Slomczynski, K. M., Jenkins, J. C., Tomescu-Dubrow, I., Kołczyńska, M., Wysmułek, I., Oleksiyenko, O., Powałko, P., Zieliński, M. W. (2017). SDR Master Box. Harvard Dataverse, V1, UNF:6:HIWud4wueVRsU8wTN+lySg== [fileUNF]. https://doi.org/10.7910/DVN/VWGF5Q

Slomczynski, K. M., Powałko, P., & Krauze, T. (2017). Non-unique Records in International Survey Projects: The Need for Extending Data Quality Control. *Survey Research Methods*, *11*(1), 1–16. https://doi.org/10.18148/srm/2017.v11i1.6557

Slomczynski, K. M., & Tomescu-Dubrow, I. (2006). Representation of Post-Communist European Countries in Cross-National Public Opinion Surveys. *Problems of Post-Communism*, *53*(4), 42–52.

Slomczynski, K. M., & Tomescu-Dubrow, I. (2018). Basic Principles of Survey Data Recycling. In T. P. Johnson, B.-E. Pennell, I. A. L. Stoop, & B. Dorer (Eds.), *Advances in Comparative Survey Methods: Multinational, Multiregional, and Multicultural Contexts (3MC)* (pp. 937–962). Wiley. https://doi.org/10.1002/9781118884997.ch43

Słomczyński, K. M., Tomescu-Dubrow, I., & Jenkins, J. C. (2016). *Democratic Values and Protest Behavior. Harmonization of Data from International Survey Projects.* Warsaw: IFiS Publishers.

Survey Research Center. (2016). *Guidelines for best practice in cross-cultural surveys*. Ann Arbor, MI: Survey Research Center, Institute for Social Research, University of Michigan. Retrieved from http://www.ccsg.isr.umich.edu/

Tarrow, S. (2011). *Power in Movement.* Social Movements and Contentious Politics, 3rd edition. Cambridge: Cambridge University Press.

UNESCO (United Nations Educational, Scientific and Cultural Organization). (2013). *UIS Methodology for Estimation of Mean Years of Schooling.* New Haven: Yale University Press. Retrieved from http://uis.unesco.org/sites/default/files/documents/uis-methodology-for-estimation-of-mean-years-of-schooling-2013-en_0.pdf

van Stekelenburg, J., & Klandermans, B. (2013). The social psychology of protest. *Current Sociology*, *61*(5–6), 886–905. https://doi.org/10.1177/0011392113479314

Verba, S., Schlozman, K. L., & Brady, H. E. (1995). *Voice and Equality: Civic Voluntarism in American Politics*. Cambridge: Harvard University Press.

Wickham, H. (2017). tidyverse: Easily Install and Load the "Tidyverse." Retrieved from https://cran.r-project.org/package=tidyverse

Wilkes, R. (2004). First Nation politics: Deprivation, resources, and participation in collective action. *Sociological Inquiry*, *74*(4), 570–589. https://doi.org/10.1111/j.1475-682X.2004.00105.x

Winship, C., & Korenman, S. D. (1997). Does Staying in School Make You Smarter?. In B. Devlin, S. E. Fienberg, D. P. Resnick, & K. Roeder (Eds.), *Intelligence, Genes and Success: Scientists Respond to the Bell Curve* (pp. 215–234). New York: Springer-Verlag.

Wolf, C., Schneider, S. L., Behr, D., & Joye, D. (2016). Harmonizing Survey Questions Between Cultures and Over Time. In *The SAGE Handbook of Survey Methodology* (pp. 502–524). SAGE. https://doi.org/10.4135/9781473957893

World Development Indicators. (2019). GDP per capita, PPP (constant 2011 international $). Retrieved June 19, 2019, from http://data.worldbank.org/indicator/NY.GDP.PCAP.PP.KD

World Values Survey. (2005). World Values Survey: Round Five (2005-2009) Questionnaire Root Version. Retrieved from www.worldvaluessurvey.org

World Values Survey. (2009). World Values Survey 1981-2008 Longitudinal Aggregate v.20090914. Madrid: World Values Survey Association. Aggregate File Producer: JD-Systems, Madrid, Spain. Retrieved from www.worldvaluessurvey.org

Appendices

Appendix A.
List of countries, project and editions included in the final analysis.

Project Edition	WVS 2	EVS 2	WVS 3	EVS 3	ASES	WVS 4	AMB 2004	ISSP 2004	AMB 2006	WVS 5	AMB 2008	EVS 4
Country												
Albania			1998			2002						2008
Algeria						2002						
Argentina	1991		1995			1999						
Armenia			1997									2008
Australia			1995							2005		
Austria		1990		1999				2004				2008
Azerbaijan			1997									2008
Bangladesh			1996			2002						
Belarus	1990		1996	2000								2008
Belgium		1990		1999								2009
Belize											2008	
Bolivia							2004		2006			
Bosnia-Herzegovina			1998			2001						2008
Brazil	1991									2006	2008	
Bulgaria		1991	1997	1999				2005		2006		2008
Burkina Faso										2007		
Canada		1990				2000		2004	2006		2007	
Chile	1990		1996			2000		2005	2006			
Colombia			1997				2004		2006	2005	2008	
Costa Rica							2004		2006		2008	

Project Edition	WVS 2	EVS 2	WVS 3	EVS 3	ASES	WVS 4	AMB 2004	ISSP 2004	AMB 2006	WVS 5	AMB 2008	EVS 4
Croatia				1999								2008
Cyprus								2004		2006		2008
Czechia	1990	1991	1998	1999				2004				2008
Denmark		1990		1999				2005				2008
Dominican Republic			1996						2006			
Ecuador									2006			
Egypt						2000						
El Salvador			1999				2004		2006			
Estonia			1996	1999								2008
Finland		1990	1996	2000				2004		2005		2009
France		1990		1999	2000			2004		2006		2008
Georgia			1996							2009		2008
Germany		1990	1997	1999	2000			2004		2006		2008
Ghana										2007		
Greece				1999	2000							2008
Guatemala							2004		2006	2005	2008	
Guyana									2006			
Haiti									2006			
Honduras							2004		2006			
Hungary		1991		1999				2004				2008
Iceland		1990		1999								2009
India	1990		1995			2001				2006		
Indonesia					2000	2001				2006		
Ireland		1990		1999	2000			2003				2008
Israel								2005				
Italy		1990		1999	2000					2005		2009
Jamaica									2006			

Project Edition	WVS 2	EVS 2	WVS 3	EVS 3	ASES	WVS 4	AMB 2004	ISSP 2004	AMB 2006	WVS 5	AMB 2008	EVS 4
Japan	1990		1995		2000			2004		2005		
Jordan						2001						
Kosovo												2008
Kyrgyzstan						2003						
Latvia			1996	1999				2004				2008
Lithuania			1997	1999								2008
Luxembourg				1999								2008
Macedonia			1998			2001						2008
Malaysia					2000					2006		
Mali										2007		
Malta		1991		1999								2008
Mexico	1990					2000	2004		2006	2005	2008	
Moldova			1996			2002				2006		2008
Montenegro			1996			2001						2008
Morocco						2001				2007		
Netherlands		1990		1999				2005		2006		2008
New Zealand			1998					2004				
Nicaragua							2004		2006			
Nigeria	1990		1995			2000						
Norway		1990	1996					2004		2007		2008
Pakistan						2001						
Panama							2004		2006			
Peru			1996			2001			2006		2008	
Philippines					2000	2001		2004				
Poland	1989	1990	1997	1999				2005				2008
Portugal		1990		1999	2000			2004				2008
Puerto Rico			1995			2001						

Project Edition	WVS 2	EVS 2	WVS 3	EVS 3	ASES	WVS 4	AMB 2004	ISSP 2004	AMB 2006	WVS 5	AMB 2008	EVS 4
Romania			1998	1999						2005		2008
Russia	1990		1995	1999				2005		2006		2008
Rwanda										2007		
Serbia			1996			2001				2006		2008
Singapore					2000							
Slovakia	1990	1991	1998	1999				2005				2008
Slovenia		1992		1999				2003		2005		2008
South Africa	1990		1996			2001		2004		2006		
South Korea	1990		1996		2000	2001		2004				
Spain		1990	1995	1999		2000		2004		2007		2008
Sweden		1990	1996	1999	2000			2004		2006		2009
Switzerland			1996					2005		2007		2008
Taiwan			1994		2000			2004		2006		
Tanzania						2001						
Thailand					2000					2007		
Trinidad & Tobago										2006		
Turkey	1990		1996	2001						2007		2009
Uganda						2001						
Ukraine			1996	1999						2006		2008
United Kingdom		1990		1999								2009
United States		1990	1995			1999		2004		2006		
Uruguay			1996					2004	2007	2006		
Venezuela			1996			2000		2004	2007			
Vietnam						2001				2006		
Zambia										2007		
Zimbabwe						2001						

methods, data, analyses | Vol. 14(1), 2020, pp. 127-154 DOI: 10.12758/mda.2020.02

The Socio-demographic Structure of the First Wave of the TwinLife Panel Study: A Comparison with the Microcensus

Volker Lang & Anita Kottwitz
Bielefeld University

Abstract

The TwinLife panel is the first longitudinal study of twin families in Germany based on a national probability sample. TwinLife has been developed to facilitate genetic sensitive research on social inequalities. The aim of this paper is to assess the usability of the TwinLife sample for such research. Therefore, first, we analyze if the social background of twins living in Germany is adequately represented in the TwinLife sample; and second, we also investigate if there are socio-demographic differences between twin and other multiple-child households in Germany which would restrict the generalizability of findings based on the TwinLife study. Specifically, we compare the distributions of key socio-demographic indicators in TwinLife with the German Microcensus using a proxy-twin and a multiple-child household sample. Our analyses show that the TwinLife sample covers the full distributions of core social inequality indicators including the lower and upper bounds, enabling researchers to use TwinLife for detailed studies of the gene-environment interplay. Furthermore, we demonstrate that (proxy-)twin and other multiple-child households in Germany are similar regarding most socio-demographic indicators. However, our analyses also indicate that participation in the first wave of the TwinLife panel was slightly selective with respect to parental education and German citizenship, especially in the younger cohorts of the study. We suggest a weighting scheme to address this selectivity.

Keywords: Twin Families, Multiple-Child Families, Family Demography, Sampling Design, Extended Twin Family Design, Germany

Studying twins reared together is a prominent research strategy to assess the influence of genetic endowment on human development (Polderman et al., 2015). By comparing monozygotic twins – who are genetically (almost) identical – with dizygotic twins – who share about half of the genes that vary between humans (like ordinary siblings), it is possible to estimate the share of variance in an outcome attributable to (additive) genetic influences (Plomin et al., 2016).[1] Nevertheless, such estimates of genetic influences are by no means a fixed quantity but strongly dependent on the development stage (i.e., the age) of the twins (Haworth et al., 2010; Turkheimer, 2000) as well as on the environmental conditions in which a genetic potential is actualized (Shanahan & Hofer, 2005; Bronfenbrenner & Ceci, 1994). A central facet of these environmental conditions is the social background (Guo & Stearns, 2002). In consequence, studying the different forms in which genetic influences depend on environments – so called gene-environment interactions and correlations – is a major focus of current behavior genetic research (Zavala et al. 2018; Tucker-Drob & Bates, 2016) as well as a topic of growing interest in the research on social inequalities (Selita & Kovas, 2019; Diewald et al., 2016; Nielsen, 2016).

However, twin samples covering a wide range of environmental conditions and development stages are needed to conduct studies on the influence of genes on social inequalities. The TwinLife panel – which is run in cooperation by research teams at Bielefeld University and Saarland University – was designed to facilitate such research and is the first longitudinal study of twin families in Germany based on a national probability sample (Mönkediek et al., 2019, Hahn et al., 2016). To assess the usability of the TwinLife sample for social stratified research on genetic influences, we address two research questions in this paper: first, is the social background of twins living in Germany captured by the TwinLife sample to facilitate genetic sensitive analyses differentiated by social background? And second, is the

1 A basic estimate of additive genetic influences is given by two times the difference between monozygotic and dizygotic twins in the correlation within twin pairs of an outcome (so called Falconer's Formula, Falconer, 1960). For a discussion of the further assumptions involved in estimating genetic influences based on twins reared together, see Stenberg (2013).

Acknowledgments

We like to thank two anonymous reviewers, Martin Diewald, and Kristina Krell for very helpful comments on an earlier version of this paper. The TwinLife project is funded by the German Research Foundation (DFG) (grant number 220286500) awarded to Martin Diewald, Rainer Riemann, and Frank M. Spinath. The TwinLife study received ethical approval from the German Psychological Association (protocol numbers: RR 11.2009 and RR 09.2013).

Direct correspondence to

Volker Lang, Bielefeld University, Department of Sociology, Project TwinLife, Postbox 100131, 33501 Bielefeld, Germany
E-mail: volker.lang@uni-bielefeld.de

social background of twin households comparable to all multiple-child households in Germany in order to support the generalizability of social stratified analyses on genetic influences?

In contrast to many other countries (e.g., The Netherlands: Ligthart et al., 2019; Sweden: Zagai et al., 2019), no twin registry is available for Germany to answer these research questions. Alternatively, we compare the TwinLife sample with two selected samples based on the German Microcensus Survey conducted by the Federal Statistical Office (Destatis, 2014a, 2014b; Lengerer et al., 2007): a proxy-twin household sample and a multiple-child household sample. Specifically, we compare parental education, household income, parental citizenship status, the composition of the households, and the population sizes of the communities of residence. In addition, we investigate maternal age at childbirth as a potential reason for differences in the distributions of these social background indicators. Thus, if the TwinLife sample is representative for twin families in Germany, we expect to see no relevant differences in the distributions of these social background indicators between the TwinLife and the Microcensus proxy-twin samples (hypothesis 1).

Moreover, since the environmental conditions in which children are reared can systematically differ between twin and other types of multiple-child families, it can be questioned if results obtained by studying twins are generalizable to a whole population. In some cases – like the age gap between siblings – such differences are undeniable. Regarding the distributions of social background indicators, differences between twin and other multiple-child families cannot be precluded. If the social backgrounds of twin and other multiple-child families in Germany are similar, we should not find any relevant differences in the distributions of the analyzed indicators between the Microcensus proxy-twin and multiple-child samples (hypothesis 2).

The paper is structured as follows. The following section describes the TwinLife and Microcensus samples as well as the indicators and methods we apply to answer our two research questions. It also contains a deeper introduction into the study design and sampling strategy of the TwinLife panel to assist researchers in using the relatively new TwinLife data. Afterwards, the results of our comparisons are presented. The final part of the article provides a conclusion.

Data and Methods

The TwinLife Panel Study

Study Design

The TwinLife study collects longitudinal data on families with monozygotic or dizygotic twin children. To exclude effects of within-twin-pair gender differences, the study includes only same-sex dizygotic twins. The base population of TwinLife consists of four birth cohorts of twins: the youngest twins, in cohort 1, were born in 2009 or 2010, the twins in cohort 2 in 2003 or 2004, the twins in cohort 3 in 1997 or 1998, and the oldest twins, in cohort 4, between 1990 and 1993. At the time of the first survey, these twins were aged around 5, 11, 17, and 23 to 24. Over the planned panel period, TwinLife covers important life course transitions ranging from school entry to the labor market entry phase, and also important life stages for meeting a partner and starting a family. The TwinLife surveys are conducted annually and survey modes alternate between face-to-face interviews at home and telephone interviews.

In addition, the TwinLife study combines this cohort-sequential design with an extended twin family design (ETFD). As part of the ETFD, the biological and, if applicable, the social parents (i.e., partners of mothers and fathers), and the sibling that is closest in age to the twins are surveyed as well as the twins themselves. Moreover, the partners of adult twins are also included. All of these family members are included in the design irrespective of whether they live in the same household as the twins or not. A family in TwinLife can therefore consist of several households, i.e., the households are nested within the families. The minimum requirement for inclusion as a valid family case in the TwinLife panel was the participation of both twins and one of the biological or social parents in the first wave.[2] A further design requirement was that the twins were raised together, i.e., lived in the same household until age 16. The family perspective of the ETFD facilitates the study of different degrees of genetic similarity which is important for detailed analysis of the manifold influences of the family environment on the development of the twins.

Sampling Strategy

The target net sample size for wave 1 of the TwinLife panel was 1,000 twin families in each of the four birth cohorts with approximately half of the families having monozygotic and the other half having same-sex dizygotic twins. To obtain a sample with these design characteristics, a national probability-based sampling procedure was implemented in two steps (Brix et al., 2017): first, a sample of 500

2 Exceptions are orphan families where there is no parent to participate. There are four families of this type in the net sample of the panel.

out of approximately 11,900 communities was drawn to generate addresses where twin families matching the design requirements resided. Potential twin families in cohorts 1 to 3 were identified by locating persons of the same sex with the same or similar birthdates registered at the same address according to the current registry of residents for the respective communities. Families in cohort 4 were also selected based on previous registries of residents containing address data prior to reported house moves. Using these previous addresses, an inquiry for the current address of the persons identified as probable twins was carried out. Second, a gross sample of 13,359 addresses out of around 19,000 addresses provided by the local registry of residents was drawn; 2,736 for cohort 1, 2,697 for cohort 2, 2,823 for cohort 3, and 5,103 for cohort 4.

Given these gross sample sizes, it was a priori obvious that the sampling design could not be proportional. Thus, each of the cohorts 1 to 3 is composed of two years of birth and cohort 4 of four years of birth. Population statistics for twin families in Germany are not available, but it is known that there are approximately 7,000 same-sex twin births each year (about 0.01 percent of all annual births, Destatis, 2013). Consequently, a design using the gross sample sizes described above and based on a cohort composed of only one year of birth would have to cover around 40 percent of the population for cohorts 1 to 3 and 75 percent for cohort 4. Using multiple-year birth cohorts reduces these shares to approximately 20 percent.

A proportional implementation of this design would necessitate conducting face-to-face interviews in around 2,500 communities which is impracticable. Three subsamples of communities were therefore selected instead: first, a proportional sample of 180 communities with 10,000 or more inhabitants was drawn according to the political community size classification for Germany (GKPOL) ("base sample"). Second, a disproportionate sample (with higher sampling probabilities for larger communities) of 60 communities with 50,000 or more inhabitants was selected to obtain the necessary coverage of the target population ("urban sample"). Third, an additional proportional sample of 260 communities with between 5,000 and 19,999 inhabitants was drawn ("rural sample").[3] The base sample consists of 5,575 addresses (41.7 percent of the gross sample), the urban sample of 6,558 addresses (49.1 percent of the gross sample), and the rural sample of 1,226 addresses (9.2 percent of the gross sample). This sampling, which is disproportional overall, leads to an overrepresentation of addresses located in urban communities in the TwinLife panel in comparison to all addresses registered in communities with 5,000 or more inhabitants (Brix et al., 2017).

3 Communities with fewer than 5,000 inhabitants are excluded by the TwinLife design. This is because, on average, only one or two twin families over all birth cohorts studied are expected to reside in a community of this size, making conducting face-to-face twin family interviews in communities like this prohibitively expensive and at the same time particularly problematic with respect to a possible re-identification.

TwinLife Sample

The gross sample of addresses described above was used for the face-to-face interviews of the TwinLife panel, wave 1. The data collection for twins born in 2009, 2003, 1997, and 1990 or 1991 was carried out between September 2014 and May 2015. For twins born in 2010, 2004, 1998, and 1992 or 1993, data collection started in September 2015 and was completed in April 2016.

Table 1 shows distributions of the gross and net samples differentiated by cohort. 10.5 percent of the addresses in the gross sample were invalid contact addresses and 4.2 percent did not comply with the requirements of the design, leaving an adjusted gross sample of 11,405 cases. In cohorts 1 to 3, around 10 percent of the cases in the adjusted gross sample were permanently absent or sick during

Table 1 Gross and net samples of TwinLife

	Cohort 1 (%)	Cohort 2 (%)	Cohort 3 (%)	Cohort 4 (%)	Total (%)
Gross sample	2,736 (100.0)	2,697 (100.0)	2,823 (100.0)	5,103 (100.0)	13,359 (100.0)
- no contact address	338 (12.4)	261 (9.7)	220 (7.8)	580 (11.4)	1,399 (10.5)
- no match with design	127 (4.6)	93 (3.4)	89 (3.2)	246 (4.8)	555 (4.2)
Adjusted gross sample	2,271 (83.0)	2,343 (86.9)	2,514 (89.1)	4,277 (83.8)	11,405 (85.4)
Adjusted gross sample	2,271 (100.0)	2,343 (100.0)	2,514 (100.0)	4,277 (100.0)	11,405 (100.0)
- absent or sick	258 (11.4)	267 (11.4)	237 (9.4)	891 (20.8)	1,653 (14.5)
- refusal	870 (38.3)	906 (38.7)	1,060 (42.2)	2,190 (51.2)	5,026 (44.1)
- family not complete	31 (1.4)	25 (1.1)	28 (1.1)	45 (1.1)	129 (1.1)
- address not used	69 (3.0)	60 (2.6)	80 (3.2)	80 (1.9)	289 (2.5)
- other reason	33 (1.5)	42 (1.8)	48 (1.9)	88 (2.1)	211 (1.9)
Net sample	1,010 (44.5)	1,043 (44.5)	1,061 (42.2)	983 (23.0)	4,097 (35.9)
Male, monozygotic	209 (20.8)	191 (18.4)	218 (20.6)	212 (21.6)	830 (20.4)
Male, dizygotic	279 (27.8)	307 (29.6)	235 (22.2)	198 (20.2)	1,019 (25.0)
Female, monozygotic	225 (22.4)	229 (22.1)	280 (26.4)	311 (31.7)	1,045 (25.6)
Female, dizygotic	291 (29.0)	309 (29.8)	326 (30.8)	259 (26.4)	1,185 (29.1)
Total	1,004 (100.0)	1,036 (100.0)	1,059 (100.0)	980 (100.0)	4,079 (100.0)

Note: The number of families used in this study declines to 4,079 compared to the net sample since in 11 families the multiples are triplets and for seven twin pairs no information about their zygosity is available.

Sources: Brix et al. (2017) and TwinLife (doi: 10.4232/1.12665), own calculations

the field phase and 40 percent refused to participate. In cohort 4, the sickness rate was twice as high and half of the sample refused participation. In 1.1 percent of the cases, it was not possible to interview all the necessary family members according to the design requirements, 2.5 percent of the addresses were not used because the target sample size had already been obtained, and 1.9 percent of the cases did not participate for other reasons.

This results in a net sample for wave 1 of 1,010 families in cohort 1, 1,043 families in cohort 2, 1,060 families in cohort 3, and 984 families in cohort 4, which closely matches the target sample size. The participation rate based on the adjusted gross sample is therefore over 40 percent in cohorts 1 to 3 and 23.0 percent in cohort 4. A total of 39 percent of the families in the net sample are part of the base sample, 51 percent are part of the urban sample, and 10.1 percent are part of the rural sample. For more information on the field process see Brix et al. (2017).

The lower part of Table 1 displays distributions by sex and zygosity of the twin pairs over the four cohorts for the net sample of the TwinLife panel.[4] There are more dizygotic than monozygotic twin pairs in cohorts 1 to 3, and in cohort 4 the share of monozygotic twin pairs is 53.3 percent. These results indicate that the probability-based sampling design used for TwinLife successfully counteracted the overrepresentation of monozygotic twins typically characterizing twin samples based on self-recruitment (i.e, two-thirds monozygotic twin pairs, with overrepresentation particularly pronounced in adult samples, Lykken et al., 1987). The findings are also in line with research showing an increase in dizygotic twining rates for OECD countries, including Germany, since the 1980s (Hoekstra et al., 2008). This is primarily because dizygotic twinning is more strongly influenced by environmental factors such as the increase in maternal age at childbirth over recent decades (Lambalk et al., 1998). Overall, the distributions demonstrate that the TwinLife sample enables genetic sensitive analyses differentiated by gender and age.

As described above, both twins, one sibling, their parents, and the partners of the adult twins are the target respondents for the interviews, irrespective of whether they live in the same household or not. Table 2 shows the composition of the families (upper part of Table 2) and the households (lower part of Table 2) interviewed in TwinLife, wave 1. Overall, the TwinLife net sample consists of 4,097 twin families living in 4,828 households. A total of 91.4 percent of these families are families with two parents.[5] However, the share of two-parent families decreases over the

4 In 50 of these families, second twin pairs exist; in 38 cases these are full siblings of the other twins, in eight cases, they are half-siblings, and in three cases, step-siblings. Moreover, one of the families has full sibling triplets in addition to the twins.

5 In 99.1 percent of the families with a mother, the mothers are the biological mothers of the twins. The share of biological fathers is 96.6 percent. In 3.8 percent of the families there are more than two parents, i.e., partners of a father or mother in addition to the biological parents.

Table 2 Family and household compositions in the net sample of TwinLife

	Cohort 1 (%)	Cohort 2 (%)	Cohort 3 (%)	Cohort 4 (%)	Total (%)
Family composition					
Mother and father, twins	431 (42.7)	337 (32.3)	350 (33.0)	290 (29.5)	1,408 (34.4)
Mother and father, twins, sibling	534 (52.9)	644 (61.7)	591 (55.7)	566 (57.6)	2,335 (57.0)
Mother or father, twins	25 (2.5)	23 (2.2)	46 (4.3)	45 (4.6)	139 (3.4)
Mother or father, twins, sibling	20 (2.0)	39 (3.7)	74 (7.0)	78 (7.9)	211 (5.2)
No parents, (sibling)[a]	0 (0)	0 (0)	0 (0)	4 (0.4)	4 (0.1)
Total	1,010 (100)	1,043 (100)	1,061 (100)	983 (100)	4,097 (100)
Household composition					
Parents, both twins, (sibling)[b]	917 (90.3)	883 (83.4)	815 (74.1)	428 (25.9)	3,043 (63.0)
Parent, both twins, (sibling)[b]	93 (9.2)	160 (15.1)	231 (21.0)	113 (6.8)	597 (12.4)
Parent(s), one twin, (sibling)[b]	0 (0)	0 (0)	22 (2.0)	184 (11.1)	206 (4.3)
Both twins, (sibling)[b]	0 (0)	0 (0)	0 (0)	84 (5.1)	84 (1.7)
One twin, (sibling)[b]	0 (0)	0 (0)	8 (0.7)	532 (32.2)	540 (11.2)
No twins	6 (0.6)	16 (1.5)	24 (2.2)	312 (18.9)	358 (7.4)
Total	1,016 (100)	1,059 (100)	1,100 (100)	1,653 (100)	4,828 (100)

[a] Orphan families; three with at least one sibling and one with no sibling.
[b] Living in a household either with or without at least one sibling.

Sources: TwinLife (doi: 10.4232/1.12665), own calculations

cohorts from 95.6 percent to 87.1 percent. In 62.2 percent of the families the twins have at least one sibling. Since parents of the earlier born twin cohorts had more time to have additional children, this share increases from 54.9 percent in cohort 1 to around 65 percent in cohorts 2 to 4. The mean number of siblings per family in families with at least one sibling is 1.6, and the maximum number of siblings is ten. Overall, the distributions indicate that TwinLife facilitates studies based on the ETFD.

The lower part of Table 2 illustrates the distribution of households in TwinLife across cohorts. As required by the study design, all of the twins in cohorts 1 and 2, and almost all of the twins in cohort 3 live together in one household. In more than 90 percent of the twin households in cohort 1, the twins live with two parents. This share drops to about 75 percent in cohort 3. For cohort 4, the share of twin

households with at least one parent is 54.1 percent. This corresponds to 43.9 percent of all households in cohort 4. A total of 76 percent of the twins from cohort 4 who had already moved out of the parental household are living without their co-twin. This represents 32.2 percent of all households in cohort 4. Further, the share of non-twin households increases from approximately 1 percent in cohorts 1 to 3 to 18.9 percent in cohort 4. These results illustrate that TwinLife captures the major shift in household structures resulting from the young adult twins starting to create their own families.[6]

The TwinLife sample for our comparisons comprises all twin households in which at least one twin resides together with at least one parent of the twins. This household definition is close to the household definition of the Microcensus (see section *The Microcensus Comparison Samples*) and retains most of the TwinLife families in the sample. This *parent-twin sample* consists of 3,640 (out of 4,828) households in TwinLife. For cohorts 1 to 3 almost all twin families and households are included in this sample. Within cohort 4, the sample covers 73.8 percent of all families and 54.1 percent of all households with twins.

The Microcensus Comparison Samples

The comparison samples we use for this study are based on the German Microcensus 2013. The Microcensus is a household survey based on a nationally representative sample of one percent (Destatis, 2014a, 2014b; Lengerer et al., 2007).[7] While the sampling of TwinLife is focused on families defined by the ETFD, the sampling design of the Microcensus is based on households, specifically persons living together at the same address sampled from the population register (Lengerer et al. 2005).

As the Microcensus survey does not collect information on whether the children living in the household are twins or not, we need to construct a suitable comparison sample to match the cohort and person composition of the TwinLife parent-twin sample described above without this information. Therefore, we define two different household samples – the multiple-child and the proxy-twin sample – based on the Microcensus. First, the *multiple-child sample* consists of one-family households with one or two parents and at least two children under the age of 25 of which at least one child – the "anchor child" – belongs to the same birth cohorts as in TwinLife. Second, the *proxy-twin sample* contains one-family households in which

6 43.4 percent of the twins in cohort 4 have a partner and 30.7 percent of these twins live in a household with their partners.

7 The 2013 Microcensus provides the most recent data currently available and thus most accurately reflects the population of 2015 – the year in which the majority of the families in the TwinLife panel was sampled.

two children of the same sex are born in the same year and live with at least one of their parents.

In view of the approximately 7,000 same-sex twin births each year (Destatis, 2013), we can expect to find around 70 proxy-twins in the 2013 Microcensus for each year of birth from circa 2000 and declining numbers for the years prior to 2000 based on the following assumptions: 1) a household sample of one percent from the population approximates a population sample of one percent; 2) there are only rare cases, other than twin births, of same-sex children in a household being born in the same year; 3) most twin children live together and with at least one parent.[8] To gain a proxy-twin sample of sufficient size for socio-demographic differentiated analyses, we use six-year birth cohorts: 2007-2012 (cohort 1), 2001-2006 (cohort 2), 1995-2000 (cohort 3), and 1989–1994 (cohort 4).

Moreover, to match the TwinLife sampling design, households in communities with fewer than 5,000 inhabitants are excluded. These represent about 16 percent of the households in the multiple-child and the proxy-twin Microcensus samples. This leaves us with 24,271 multiple-child and 1,039 proxy-twin households for our analysis.

Indicators

With respect to the social structural indicators used for the analysis, we compare household structures, the size of the communities where the household is located, German citizenship status on the household level, highest education of parents in the household, and also monthly net equivalent household income in euros. To assess the potential use of the TwinLife study for multidimensional analysis of social structural (dis-)advantage, we also compare the bivariate distributions of highest education in the household by monthly net equivalent household income. Moreover, we contrast maternal age at birth of the twins or the anchor child as a potential reason for social structural differences between the samples since giving birth later in life could be correlated with higher educational degrees or higher earnings.

The size of the community where the household is located is categorized based on the German community size classification (GKPOL). German citizenship is used as a proxy for migration background since the alternative indicators for migration background available in TwinLife and the Microcensus are not comparable. We assign German citizenship status on the household level if both parents

8 There are rarely any women who give birth to two children within the same calendar year. However, the Microcensus does not differentiate between biological and stepchildren. Thus, there might be a negligible number of cases which are spuriously considered as twin families. These might be foster or blended families with same-sex children born in the same year.

have German citizenship. The highest education within the household is based on the International Standard Classification of Education (ISCED) 1997 (Schneider, 2008). The individual-level information on parents' education is used to calculate the highest obtained degree on the household level. The ISCED is coded as an ordered categorical variable with "no educational degree" (1) as the lowest and "Ph.D. degree" (6) as the highest category. Information on monthly net income is surveyed on the household level. To make the household incomes comparable across different household structures, an equivalence weight according to the new OECD scheme (OECD 2011) and an adjustment for inflation dividing the nominal income by the Consumer Price Index for Germany using 2015 as base year are applied.

Methods

To assess whether distributions of the social background indicators differ between the samples, we construct categorical variables based on these indicators and calculate the proportion of each category for the distributions of these categorical variables. In addition, we perform z-tests on equality of proportions between samples using the 95% confidence level and report their statistical significance for the substantial differences discussed in this paper. Cell-specific case numbers in the Microcensus proxy-twin sample are too small to show detailed distributions for highest ISCED in households and net equivalent monthly household income. Thus, we present ISCED levels 5a and 6 versus all lower levels and household's median income. For maternal age at childbirth, we compare the means.

To account for missing values in education, citizenship and monthly net household income in the TwinLife sample, we set up a multiple imputation model on the household level.[9] We impute 20 values for each missing observation using multiple imputation with chained equations (van Buuren et al., 2006), a method which iterates over a sequence of univariate imputation models for each variable. For the univariate imputation models, we use predictive mean matching with ten nearest neighbors in case of continuous variables and logistic or ordered logistic regressions in case of categorical variables.[10] The procedure assumes that the data is missing at random conditional on the predictors used. To preferably ensure that

9 Information is missing on ISCED for 4.5 percent of the mothers and 22.9 percent of the fathers, on German citizenship status for 4 percent of the mothers and 22.6 percent of the fathers, and on monthly net household income for 12.2 percent of the households.

10 The values presented in the descriptions are calculated as the mean of imputations in case of continuous and as the mode of imputations in case of categorical variables.

this assumption is met, we use a comprehensive set of predictors.[11] We assess the influence of the imputation procedure on the distributions of the social structural indicators compared. Here, we find slight increases in the lower categories of the indicators (typically about 2 percent) and converse declines in the upper categories. However, there are only minor differences between imputed and non-imputed estimates. Thus, in the following results section, we refrain from presenting non-imputed in addition to imputed results for reasons of clarity and brevity.

Results

Comparisons of the Social Background Indicators

In this section, we present the results of the comparisons of the distributions of the social background indicators in the TwinLife parent-child, the Microcensus proxy-twin, and the Microcensus multiple-child sample.

Household Structure

Table 3 shows the household structures in the TwinLife parent-twin sample in contrast to the two Microcensus comparison samples. The number of children living in a household with both parents differs in the Microcensus multiple-child sample compared to the TwinLife parent-twin and the Microcensus proxy-twin samples.

While there are 58.9 percent of households with two children and both parents in the former sample, this share is approximately 40 percent in the latter two. This difference is plausible since potential parents often plan to have two children but if the second birth is a twin birth, they have three children (Ruckdeschel, 2007). The share of single-parent households is about 16 percent in all three samples. Overall, these results indicate that the main difference in the composition of twin and non-twin multiple-child households is the higher prevalence of households with two children in the latter group. In addition, the findings confirm that the probability-based sampling procedure used for TwinLife was appropriate in this regard since the household structures in the TwinLife parent-twin and the Microcensus proxy-twin samples are similar.

11 We use all imputed variables as well as information on the years of birth, migration background, cognitive test scores, monthly gross income and weekly working hours of mothers and fathers, household structure, and community size as predictors.

Table 3 Household structures in the TwinLife and Microcensus comparison samples

	Cohort 1 (%)	Cohort 2 (%)	Cohort 3 (%)	Cohort 4 (%)	Total (%)
TwinLife parent-twin sample					
Couples, twin(s)	428 (42.4)	355 (34.0)	401 (38.3)	259 (47.9)	1443 (39.6)
Couples, twin(s), sibling	489 (48.4)	528 (50.6)	414 (39.6)	169 (31.2)	1600 (44.0)
Single parent, twin(s)	50 (5.0)	80 (7.7)	149 (14.2)	76 (14)	355 (9.8)
Single parent, twin(s), sibling	43 (4.3)	80 (7.7)	82 (7.8)	37 (6.8)	242 (6.6)
Total	1,010 (100.0)	1,043 (100.0)	1,046 (100.0)	541 (100.0)	3,640 (100.0)
Microcensus multiple-child sample					
Couples, 2 children	3,680 (61.1)	3,523 (55.6)	3,531 (55.7)	3,558 (63.9)	14,292 (58.9)
Couples, 3 or more children	1,713 (28.5)	1,774 (28.0)	1,544 (24.3)	948 (17.0)	5,979 (24.6)
Single parent, 2 children	426 (7.1)	732 (11.5)	958 (15.1)	924 (16.6)	3,040 (12.5)
Single parent, 3+ children	199 (3.3)	310 (4.9)	309 (4.9)	142 (2.5)	960 (4.0)
Total	6,018 (100.0)	6,339 (100.0)	6,342 (100.0)	5,572 (100.0)	24,271 (100)
Microcensus proxy-twin sample					
Couples, 2 children	139 (46.8)	82 (28.3)	99 (33.2)	70 (45.5)	390 (37.5)
Couples, 3 or more children	122 (41.1)	149 (51.4)	139 (46.6)	48 (31.2)	458 (44.1)
Single parent, 2 children	20 (6.7)	34 (11.7)	30 (10.1)	27 (17.5)	111 (10.7)
Single parent, 3+ children	16 (5.4)	25 (8.6)	30 (10.1)	9 (5.8)	80 (7.7)
Total	297 (100.0)	290 (100.0)	298 (100.0)	154 (100.0)	1,039 (100.0)

Sources: TwinLife (doi: 10.4232/1.12665) and Microcensus 2013, own calculations

Community Size

Table 4 reports shares of households by community size across the three samples. Around two-thirds of the TwinLife households are located in communities with 50,000 or more inhabitants while this share is around 40 percent in the Microcensus samples.

This difference is statistically significant and mainly attributable to the oversampling of urban communities in TwinLife which was implemented to obtain the necessary coverage of the twin family target population (see sub-section *Sampling*

Strategy). However, if we exclude the oversampled urban population, the distributions of the TwinLife and Microcensus samples are roughly comparable. The group of TwinLife households in communities with 500,000 or more inhabitants is around four percentage points larger than the Microcensus samples, and the share of households in communities with 100,000 to 499,999 inhabitants is approximately six percentage points smaller in the TwinLife sample than in the Microcensus samples. The latter of these two differences is statistically significant. Regarding the Microcensus proxy-twin and multi-child samples, there are no considerable differences in shares of households by community size between the samples.

Table 4 Households by community size in percent

	Cohort 1	Cohort 2	Cohort 3	Cohort 4	Total
TwinLife parent-twin sample					
5,000–19,999 (in %)	18.4	18.9	19.2	21.8	19.3
20,000–49,999 (in %)	10.5	13.3	10.9	14.1	12.0
50,000–99,999 (in %)	18.0	16.1	15.2	16.1	16.4
100,000–499,999 (in %)	21.9	21.1	22.6	20.5	21.7
> 500,000 (in %)	31.2	30.6	32.1	27.5	30.7
TwinLife, without urban sample					
5,000–19,999 (in %)	38.4	37.4	39.0	37.0	38.1
20,000–49,999 (in %)	20.1	25.0	21.1	23.1	22.3
50,000–99,999 (in %)	10.1	9.3	8.8	10.5	9.5
100,000–499,999 (in %)	11.7	11.0	11.0	7.5	10.6
> 500,000 (in %)	19.7	17.3	20.1	22.0	19.5
Microcensus multiple-child sample					
5,000–19,999 (in %)	31.3	33.9	35.5	35.6	34.1
20,000–49,999 (in %)	22.1	23.8	23.5	24.1	23.4
50,000–99,999 (in %)	10.6	10.1	11.0	11.4	10.8
100,000–499,999 (in %)	17.4	16.2	15.9	15.6	16.3
> 500,000 (in %)	18.7	15.9	14.1	13.3	15.5
Microcensus proxy-twin sample					
5,000–19,999 (in %)	26.6	33.8	35.6	31.8	32.0
20,000–49,999 (in %)	22.9	22.1	20.1	23.4	21.9
50,000–99,999 (in %)	11.5	11.0	12.1	13.6	11.8
100,000–499,999 (in %)	18.2	17.9	15.1	18.8	17.3
> 500,000 (in %)	20.9	15.2	17.1	12.3	16.9

Sources: TwinLife (doi: 10.4232/1.12665) and Microcensus 2013, own calculations

Parental Citizenship Status

Table 5 contrasts the shares of households with German citizenship across the samples. Overall, this share is 84.7 percent in the TwinLife sample while the corresponding shares are around 80 percent in the Microcensus samples. The share is constant across cohorts in the TwinLife sample while it declines in the Microcensus samples from about 85 percent in the older cohorts to about 75 percent in the younger cohorts. Consequently, there are around five to ten percentage points more households with German citizenship in the TwinLife sample for cohorts 1 and 2 and these differences are statistically significant. The shares of households with German citizenship in the Microcensus proxy-twin and multiple-child samples are similar.

Table 5 Households by German citizenship

	Cohort 1	Cohort 2	Cohort 3	Cohort 4	Total
TwinLife parent-twin sample					
German citizenship (in %)	85.5	85.0	84.1	83.7	84.7
No German citizenship (in %)	14.5	15.0	15.9	16.3	15.3
Microcensus multiple-child sample					
German citizenship (in %)	74.1	77.9	82.3	81.7	79.0
No German citizenship (in %)	25.9	22.1	17.7	18.3	21.1
Microcensus proxy-twin sample					
German citizenship (in %)	75.8	76.9	85.9	85.7	80.5
No German citizenship (in %)	24.2	23.1	14.1	14.3	19.5

Sources: TwinLife (doi: 10.4232/1.12665) and Microcensus 2013, own calculations

Parental Education

Table 6 describes the distributions of highest educational level in the households for the TwinLife parent-twin and the Microcensus multiple-child samples based on the ISCED. We observe that the TwinLife sample covers the full distribution of educational levels. The lower tail (ISCED 1 and 2) encompasses around 5 percent of the cases. The results indicate that there are more households with a university education (ISCED 5a and 6) and fewer with medium or low education (ISCED 1 to 3) in TwinLife than the Microcensus multiple-child sample, particularly in the younger cohorts.

To analyze potential reasons for these differences, the lower part of Table 6 shows the shares of university educated households compared to all other house-

Table 6 Highest educational level (based on ISCED) in household

	Cohort 1	Cohort 2	Cohort 3	Cohort 4	Total
TwinLife parent-twin sample					
ISCED 1 (in %)	0.6	0.8	1.1	1.5	0.9
ISCED 2 (in %)	4.9	4.0	3.6	5.2	4.3
ISCED 3a, b, c (in %)	25.2	27.0	33.5	37.9	30.0
ISCED 4a, b (in %)	9.8	7.2	8.0	8.1	8.3
ISCED 5b (in %)	10.8	13.2	12.8	16.5	12.9
ISCED 5a (in %)	41.4	42.7	36.0	27.9	38.2
ISCED 6 (in %)	7.4	5.1	5.0	3.0	5.4
Microcensus multiple-child sample					
ISCED 1 (in %)	3.4	3.7	3.6	4.2	3.7
ISCED 2 (in %)	9.9	8.6	7.9	10.2	9.1
ISCED 3a, b, c (in %)	34.1	36.8	39.3	43.5	38.3
ISCED 4a, b (in %)	9.4	9.2	9.2	7.8	8.9
ISCED 5b (in %)	11.5	12.7	14.7	14.0	13.2
ISCED 5a (in %)	27.8	25.5	22.7	18.4	23.7
ISCED 6 (in %)	4.0	3.4	2.6	2.0	3.0
TwinLife parent-twin sample					
ISCED 1, 2, 3, 4, or 5b (in %)	51.3	52.3	59.1	69.2	46.5
ISCED 5a or 6 (in %)	48.7	47.7	40.9	30.8	43.5
Microcensus multiple-child sample					
ISCED 1, 2, 3, 4, or 5b (in %)	68.2	71.1	74.7	79.6	73.3
ISCED 5a or 6 (in %)	31.8	28.9	25.3	20.4	26.7
Microcensus proxy-twin sample					
ISCED 1, 2, 3, 4, or 5b (in %)	43.6	74.4	76.4	77.8	72.4
ISCED 5a or 6 (in %)	36.4	25.6	23.6	22.2	27.6
TwinLife, without urban sample					
ISCED 1, 2, 3, 4, or 5b (in %)	52.9	57.1	60.3	69.5	58.9
ISCED 5a or 6 (in %)	47.1	42.9	39.7	30.5	41.1
TwinLife, only German citizenship					
ISCED 1, 2, 3, 4, or 5b (in %)	48.7	48.8	55.8	66.1	53.3
ISCED 5a or 6 (in %)	51.3	51.2	44.2	33.9	46.7
Microcensus multiple-child sample, only German citizenship					
ISCED 1, 2, 3, 4, or 5b (in %)	65.1	65.8	72.8	77.3	70.8
ISCED 5a or 6 (in %)	34.9	34.2	27.2	22.7	29.2

Note: Cell-specific case numbers in the Microcensus proxy-twin sample are too small to present detailed distributions for highest ISCED in households.

Sources: TwinLife (doi: 10.4232/1.12665) and Microcensus 2013, own calculations

holds. Overall, the share of university educated households is 43.5 percent in the TwinLife sample while it is around 27 percent in the Microcensus samples. In cohort 4 the difference is around ten percentage points between the samples while it is between 15 and 20 percentage points in cohorts 1 to 3. All of these differences are statistically significant. The differences in younger cohorts decline slightly if we restrict the samples to households with German citizenship to account for the higher shares of these households in TwinLife.[12] The shares of households with a university education in the Microcensus proxy-twin and multiple-child samples are approximately the same.

Household Income

Table 7 reports the distributions of monthly net equivalent household incomes for the TwinLife and Microcensus samples. It can be shown that the TwinLife sample covers the full income distribution. Across all cohorts, around 20 percent of the households have an adjusted income of less than €1,000 per month, around 53 percent have between €1,000 and €2,000 per month, around 20 percent have between €2,000 and €3,000 per month, and approximately 7 percent have more than €3,000 per month.

These shares are roughly comparable to the Microcensus samples where the share of households with less than €1,000 per month is slightly higher and the share with between €2,000 and €3,000 per month is slightly lower. For these two income categories the differences between the TwinLife sample and the Microcensus samples are statistically significant. Overall, the median monthly net equivalent household income in the TwinLife sample is €1,528 while it is around €150 less in the Microcensus samples. Differentiated by cohort, these differences between monthly median incomes are approximately €100 in cohorts 3 and 4 and around €200 in cohorts 1 and 2. Restricting the TwinLife and Microcensus samples to households with German citizenship or excluding the TwinLife urban sample does not account for the differences observed. Conditional on parental education the household income medians are similar in the TwinLife and the Microcensus samples. This finding indicates that the differences in household income between the samples are mostly a consequence of the selective participation in TwinLife with respect to parental education (see sub-section *Parental Education*).

12 Excluding the urban sample of TwinLife to address the oversampling of urban households in TwinLife (see sub-section *Parental Education*) does not change the differences observed in the shares of university educated households between TwinLife and the Microcensus samples to a relevant degree.

Table 7 Monthly net equivalent household income

	Cohort 1	Cohort 2	Cohort 3	Cohort 4	Total
TwinLife parent-twin sample					
Household income in € (median)	1,618	1,574	1,403	1,610	1,528
Household income in € (in %):					
< €1,000	18.3	18.6	25.5	17.4	20.3
€1,000 to < €2,000	50.9	52.4	54.4	51.9	52.5
€2,000 to < €3,000	23.3	20.9	15.1	23.8	20.3
≥ €3,000	7.5	8.2	5.0	6.8	6.9
Microcensus multiple child sample					
Household income in € (median)	1,324	1,373	1,376	1,537	1,375
Household income in € (in %):					
< €1,000	26.6	23.9	25.8	19.8	24.1
€1,000 to < €2,000	50.8	50.1	55.0	56.6	53.0
€2,000 to < €3,000	16.2	18.3	14.2	18.7	16.8
≥ €3,000	6.4	7.7	4.9	5.0	6.0
Microcensus proxy-twin sample					
Household income in € (median)	1,433	1,285	1,303	1,537	1,373
Median income in € in subsamples					
TwinLife, without urban sample	1,574	1,549	1,405	1,612	1,520
TwinLife, only German citizenship	1,670	1,670	1,499	1,733	1,664
Microcensus multiple child sample, only German citizenship	1,469	1,478	1,433	1,601	1,495

Note: Cell-specific case numbers in the Microcensus proxy-twin sample are too small to present detailed distributions for net equivalent monthly household income.

Sources: TwinLife (doi: 10.4232/1.12665) and Microcensus 2013, own calculations

Parental Education and Household Income Combined

In Table 8 the monthly net equivalent household income distributions are further differentiated by the highest educational status in the households based on ISCED in order to assess the TwinLife studies potential for multidimensional analysis of social structural (dis-)advantage. The parts of this two-dimensional social structural distribution covered in the Microcensus multiple-child sample are also represented in the TwinLife parent-twin sample indicating that the latter can be used for related multidimensional analysis. Further, the distributions are also roughly comparable; the shares of households with a university education (ISCED 5a or 6) and an adjusted income of between €1,000 and €3,000 are larger in the TwinLife parent-twin sample, while those with medium education (ISCED 3) and an income

of between €1,000 and €2,000 and also those with low education (ISCED 1 or 2) and an income of less than €1,000 are lower.

Table 8 Highest educational level (ISCED) by net equivalent income in households

Monthly net equivalent household income	TwinLife parent-twin sample				Microcensus multiple child sample			
	Highest educational level (based on ISCED) in household in % (cell percentages)							
	1, 2	3a, b, c	4a, b, 5b	5a, 6	1, 2	3a, b, c	4a, b, 5b	5a, 6
Cohort 1								
< €1,000	3.8	7.8	2.2	4.6	10.0	11.6	2.8	2.2
€1,000 to < €2,000	1.7	16.0	14.1	19.1	3.1	20.9	13.8	13.0
€2,000 to < €3,000	0	1.1	4.0	18.2	0.1[a]	1.8[a]	3.4	11.2
≥ €3,000	0	0.2	0.4	6.9			0.6	5.4
Cohort 2								
< €1,000	2.6	8.6	3.3	4.1	8.3	10.7	3.1	1.7
€1,000 to < €2,000	2.2	16.4	13.7	20.0	4.1	23.1	13.4	9.6
€2,000 to < €3,000	0	1.7	2.8	16.4	0.2[a]	2.9[a]	4.2	11.6
≥ €3,000	0	0.3	0.7	7.2			1.1	6.0
Cohort 3								
< 1,000	3.7	12.5	3.7	5.5	7.9	12.7	3.2	2.0
€1,000 to < €2,000	1.0	18.9	14.4	20.1	3.4	24.2	16.2	11.2
€2,000 to < €3,000	0	1.7	2.3	11.1	0.2[a]	2.5[a]	3.4	8.5
≥ €3,000	0	0.3	0.4	4.3			0.9	3.6
Cohort 4								
< €1,000	2.8	7.4	3.0	4.3	7.3	8.9	2.3	1.2
€1,000 to < €2,000	3.1	21.8	15.0	12.0	6.8	28.0	13.1	8.8
€2,000 to < €3,000	0.7	7.2	6.3	9.6	0.5[a]	6.8[a]	4.9	7.2
≥ €3,000	0	1.5	0.4	5.0			1.1	3.2

[a] Due to small sample sizes, the shares of the categories €2,000 to < €3,000 and ≥ €3,000 are aggregated for ISCED 1, 2 and ISCED 3a, b, c.

Sources: TwinLife (doi: 10.4232/1.12665) and Microcensus 2013, own calculations

Maternal Age at Birth

Finally, we compare the mean values of maternal age at the birth of the twins or the anchor child for the TwinLife and Microcensus samples. This value is approximately 31 years in all samples and the differences between samples are statistically not significant. It increases from around 30 years in cohort 4 to about 32 years in cohort 1 which is accompanied by an increase in the share of mothers aged 35 or older at childbirth (from around 15 to 30 percent). The changes are less pronounced in the Microcensus multiple-child sample. Overall, there are no indications of differences in maternal age at childbirth which could be responsible for the social structural differences observed.

Limitations

With respect to the comparisons conducted in this study, the main limitation is the lack of a twin registry for Germany. Thus, we had to use a proxy-twin sample which is based on a one percent general population sample. As a result, the size of the proxy-twin sample is small. Moreover, we cannot conduct comparative analyses differentiating between monozygotic and dizygotic twins since there is no information on zygosity available for the proxy-twins. Nevertheless, our comparison samples are based on the Microcensus, a survey of high quality standards, particularly regarding representativity (Lengerer et al., 2007). Therefore, the Microcensus is the best dataset available for conducting a study on the generalizability of socio-structural differentiated analyses of twins in Germany.

The central limitation our study found with respect to using TwinLife for such analyses is the slight selectivity of the TwinLife sample with respect to parental education and German citizenship. Partly, the underrepresentation of families without German citizenship is due to conducting the study only in German and restricting the sampling to families with sufficient proficiency of the German language (Brix et al., 2017). The underrepresentation of respondents with migration background – often corresponding with having no German citizenship – can commonly be addressed using specialized sampling strategies (Brücker et al., 2014; Schupp & Wagner, 1995). However, TwinLife did not have funding for instruments in additional languages or an additional migration sample. A potential reason for the selectivity regarding parental education is the demanding questionnaire program for the first wave of TwinLife, particularly for the children aged around 5 at the time of the survey in cohort 1. To ensure panel stability, plans had already been made to shorten the survey for future TwinLife waves prior to the first wave and the program has been further reduced given the results of this study.[13]

13 The expected workload on the family level for the second wave was reduced from around 180 minutes in the first wave to around 120 minutes.

Selectivity Correction

To address the selective participation in TwinLife with regard to parental education and German citizenship (see sub-sections *Parental Citizenship Status* and *Parental Education*), we suggest conducting additional analyses using a cohort-specific weighting scheme based on the distribution of highest education in the households by German citizenship in the Microcensus multiple-child sample (see Appendix A). Since household income levels conditional on parental education are similar in both samples (see sub-section *Household Income*), we consider the differential incomes a consequence of the differences in education. In consequence, we did not include household income as additional indicator in our proposed weighting scheme. In principle, using such a weighting scheme for TwinLife is justified by the social structural similarity between (proxy-)twin and multiple-child households in Germany found in this study.

Conclusion

In this paper, we addressed two research questions regarding the generalizability of research on the gene-environment interplay utilizing the TwinLife data: first, we assessed the usability of the TwinLife sample for social stratified analyses of genetic influences; and second, we analyzed whether the social background of twin households in Germany is comparable to the whole population of multiple-child households. Furthermore, we introduced the design and sampling strategy of TwinLife to assist researcher in using the TwinLife panel for their research.

Social Stratified Genetic Sensitive Analyses using TwinLife

Addressing our first research question, our comparison shows larger shares of urban households in TwinLife due to the oversampling of populous communities that was necessary to achieve the target sample size. Furthermore, the share of households with migration background – indicated by no German citizenship – is approximately five to ten percentage points smaller in the younger cohorts of the TwinLife compared to the Microcensus samples. Moreover, we show that the probability-based sampling of the TwinLife study was successful in counteracting the overrepresentation of monozygotic twins typical of twin samples based on self-recruitment (Lykken et al., 1987).

Looking at the core socio-economic indicators – parental education and income – our results show that the TwinLife sample covers the full distributions including the lower and upper bounds. With regard to parental education, we found around 15 percentage points more university educated households in the Twin-

Life sample, particularly in the younger cohorts. The smaller share of households with no German citizenship in TwinLife can explain some of the differences in the shares of university educated households between the samples. For the monthly net equivalent household income, we found that median values were around €200 higher for the younger TwinLife cohorts and that the corresponding values were around €100 higher in the older cohorts. Additional analyses showed that the oversampling of urban communities in TwinLife cannot account for these differences.

In sum, our findings indicate that participation in TwinLife was, to some degree, selective with respect to parental education and German citizenship, specifically in the younger cohorts. We proposed a weighting scheme to address this selectivity. However, since the TwinLife sample covers the whole distributions of the social background indicators, this selectivity does not restrict the usability of the TwinLife sample for social stratified analyses of genetic influences. In principle, TwinLife can be used for multidimensional analyses of genetic influences on social inequalities based on an ETFD.

Social Background Differences between Twin and Multiple-child Households

Regarding our second research question, our analyses show that (proxy-)twin and multiple-child households in Germany have comparable distributions for many socio-demographic indicators such as community size, parental citizenship status, parental education, household income, and maternal age at birth of the twins or anchor children. The only difference we found between twin and multiple-child households is the higher prevalence of households with two children in the latter group. This difference can be explained by parents often planning to have two children (Ruckdeschel, 2007).

The absence of relevant differences in the distributions of social background indicators between twin and other multiple-child households is important for TwinLife, since it would otherwise be impossible to capture the full range of social structural variation using a twin-based sampling strategy. Moreover, this is also beneficial for generalizing inferences of social structural influences based on the TwinLife sample to the corresponding population at large. If different outcomes in twin and other multiple-child families are not a consequence of different social structural distributions, these varying outcomes indicate different inequality-generating processes for twin and non-twin families. Therefore, if a researcher has reason to assume that there are no different inequality-generating processes for twin and non-twin families, findings based on the TwinLife data can be generalized to all multiple-child families in Germany.

References

Brix, J., Pupeter, M., Rysina, A., Steinacker, G., Schneekloth, U., Baier, T., . . . Spinath, F. M. (2017). *A longitudinal twin family study of the life course and individual development (TWINLIFE): Data collection and instruments of wave 1 face-to-face interviews. TwinLife Technical Report Series: Vol. 5.* Bielefeld / Saarbrücken. Retrieved from https://pub.uni-bielefeld.de/record/2914569

Bronfenbrenner, U., & Ceci, S. J. (1994). Nature-nuture reconceptualized in developmental perspective: A bioecological model. *Psychological Review, 101(4)*, p. 568–586. https://doi.org/10.1037/0033-295X.101.4.568

Brücker, H., Kroh, M., Bartsch, S., Goebel, J., Kühne, S., Liebau, E.,. . . Schupp, J. (2014). *The new IAB-SOEP migration sample: An introduction into the methodology and the contents. SOEP survey papers Series C, Data documentations: Vol. 216*. Berlin: DIW. Retrieved from http://panel.gsoep.de/soep-docs/surveypapers/diw_ssp0216.pdf

Destatis. (2013). *Natürliche Bevölkerungsbewegung.* Wiesbaden.

Destatis. (2014a). *Bevölkerung und Erwerbstätigkeit – Haushalte und Familien: Ergebnisse des Mikrozensus 2013.* Wiesbaden. https://doi.org/10.1007/978-3-658-11490-9_16

Destatis. (2014b). *Mikrozensus 2013: Qualitätsbericht*. Wiesbaden.

Diewald, M., Baier, T., Schulz, W., Schunck, R. (2016). Status attainment and social mobility. How can genetics contribute to an understanding of their causes? In K. Hank & M. Kreyenfeld (eds), *Social Demography Forschung an der Schnittstelle von Soziologie und Demografie. Kölner Zeitschrift für Soziologie und Sozialpsychologie* (Sonderheft 55/2015, p. 371–395), Wiesbaden. https://doi.org/10.1007/978-3-658-11490-9_16

Falconer, D. S. (1960). *Introduction to quantitative genetics. (1st ed.)* New York: Ronald Press Co.

Guo, G., & Stearns, E. (2000). The social influences on the realization of genetic potential for intellectual development, *Social Forces, 80(3)*, p. 881–910. https://doi.org/10.1353/sof.2002.0007

Hahn, E., Gottschling, J., Bleidorn, W., Kandler, C., Spengler, M., Kornadt, A. E., . . . Spinath, F. M. (2016). What drives the development of social inequality over the life course?: The German TwinLife Study. *Twin Research and Human Genetics, 19(6)*, p. 659–672. https://doi.org/10.1017/thg.2016.76

Haworth, C. M. A., Wright, M. J., Luciano, M., Martin, N. G., de Geus, E. J. C., van Beijsterveldt, C. E. M., . . . Plomin, R. (2010). The heritability of general cognitive ability increases linearly from childhood to young adulthood. *Molecular Psychiatry, 15*, p. 1112–1120. https://doi.org/10.1038/mp.2009.55

Hoekstra, C., Zhao, Z. Z., Lambalk, C. B., Willemsen, G., Martin, N. G., Boomsma, D. I., & Montgomery, G. W. (2008). Dizygotic twinning. *Human Reproduction Update, 14(1)*, p. 37–47. https://doi.org/10.1093/humupd/dmm036

Lambalk, C. B., De Koning, C. H., & Braat, D. D. M. (1998). The endocrinology of dizygotic twinning in the human, *Molecular and Cellular Endocrinology, 145(1-2)*, p. 97–102. https://doi.org/10.1016/S0303-7207(98)00175-0

Lengerer, A., Bohr, J., & Janßen, A. (2005). *Households, families and ways of life in the microcensus: concepts and stylizations (ZUMA-Arbeitsbericht).* Mannheim.

Lengerer, A., Janßen, A., & Bohr, J. (2007). Possibilities for family research with the German Microcensus. *Journal of Family Research, 19(2)*, p. 186–209. https://nbn-resolving.org/urn:nbn:de:0168-ssoar-58103

Ligthart, L., Van Beijsterveldt, C., Kevenaar, S., De Zeeuw, E., Van Bergen, E., Bruins, S., . . . Boomsma, D. (2019). The Netherlands Twin Register: Longitudinal research based on twin and twin-family designs. *Twin Research and Human Genetics, online first*, p. 1–14. https://doi.org/10.1017/thg.2019.93

Lykken, D. T., McGue, M., & Tellegen, A. (1987). Recruitment bias in twin research: The rule of two-thirds reconsidered. *Behavior Genetics, 17(4)*, p. 343–362. http://dx.doi.org/10.1007/BF01068136

Mönkediek, B., Lang, V., Weigel, L., Baum, M., Eifler, E., Hahn, E., . . . Spinath, F. (2019). The German Twin Family Panel (TwinLife). *Twin Research and Human Genetics, online first*, p. 1–8. https://doi.org/10.1017/thg.2019.63

Nielsen, F. (2016). The status-achievement process: Insights from genetics. *Frontiers in Sociology, 1(9)*, p. 1–15. https://doi.org/10.3389/fsoc.2016.00009

OECD. (2011). *What Are Equivalence Scales?: OECD Project on Income Distribution and Poverty.* Retrieved from http://www.oecd.org/eco/growth/OECD-Note-EquivalenceScales.pdf, 25 April 2017

Plomin, R., DeFries, J. C., Knopik, V. S., & Neiderhiser, J. (2016). *Behavioral genetics. (7th ed.)* New York: Worth Publishers.

Polderman, T. J. C., Benyamin, B., de Leeuw, C. A., Sullivan, P. F., van Bochoven, A., Visscher, P. M., & Posthuma, D. (2015). Meta-analysis of the heritability of human traits based on fifty years of twin studies. *Nature Genetics 47(7)*, p. 702–709. https://doi.org/10.1038/ng.3285

Ruckdeschel, K. (2007). Fertility intentions of childless persons. *Journal of Family Research, 19(2)*, p. 210–230. https://nbn-resolving.org/urn:nbn:de:0168-ssoar-58113

Schneider, S. L. (2008). Applying the ISCED-97 to the German educational qualifications. In S. L. Schneider (Ed.), *The International Standard Classification of Education (ISCED-97). An evaluation of content and criterion validity for 15 European countries* (p. 76–102). Mannheim.

Schupp, J., & Wagner, G. G. (1995). Die Zuwanderer-Stichprobe des Sozio-ökonomischen Panels (SOEP). *Vierteljahrshefte zur Wirtschaftsforschung, 64(1)*, p. 16–25. https://www.econstor.eu/bitstream/10419/141079/1/vjh_v64_i01_pp016-025.pdf

Selita, F., & Kovas, Y. (2019). Genes and Gini: What inequality means for heritability. *Journal of Biosocial Science, 51(1)*, p. 18–47. https://doi.org/10.1017/S0021932017000645

Shanahan, M. J., & Hofer, S. M. (2005). Social context in gene-environment interactions: Retrospect and prospect, *The Journals of Gerontology: Series B, 60(1)*, p. 65–76. https://doi.org/10.1093/geronb/60.Special_Issue_1.65

Strenberg, A. (2013). Interpreting estimates of heritability – A note on the twin decomposition, *Economics and Human Biology, 11(2)*, p. 201–205. https://doi.org/10.1016/j.ehb.2012.05.002

Tucker-Drob, E. M., & Bates, T.C. (2016) Large cross-national differences in gene × socioeconomic status interaction on intelligence. *Psychological Science 27(2)*, p. 138–149. https://doi.org/10.1177/0956797615612727

Turkheimer, E. (2000). Three laws of behavior genetics and what they mean. *Current Directions in Psychological Science*, 9*(5),* p. 160–164. https://doi.org/10.1111/1467-8721.00084

van Buuren, S., Brand, J. P. L., Groothuis-Oudshoorn, C. G. M., & Rubin, D. B. (2006). Fully conditional specification in multivariate imputation. *Journal of Statistical Computation and Simulation, 76(12)*, p. 1049–1064. https://doi.org/10.1080/10629360600810434

Zagai, U., Lichtenstein, P., Pedersen, N., & Magnusson, P. (2019). The Swedish Twin Registry: Content and management as a research infrastructure. *Twin Research and Human Genetics, online first*, p. 1–9. https://doi.org/10.1017/thg.2019.99

Zavala, C., Beam, C. R., Finch, B. K., Gatz, M., Johnson, W., Kremen, W. S., . . . Reynolds, C. A. (2018). Attained SES as a moderator of adult cognitive performance: Testing gene–environment interaction in various cognitive domains. *Developmental Psychology, 54(12)*, p. 2356–2370. https://doi.org/10.1037/dev0000576

Appendix A

Selectivity correction weighting scheme based on the Microcensus

This appendix contains instructions for constructing a weighting scheme matching the cohort specific highest ISCED by German citizenship distribution of parents on the household level for TwinLife analysis samples with the Microcensus multiple-child sample. The aim of the proposed weighting scheme is to address the selectivity of the TwinLife sample regarding parental education and German citizenship status, particularly in the younger cohorts. We advise using it as a robustness check, i.e., to assess discrepancies in the results between analyses conducted with and without the weighting scheme. Comparable results in both analyses indicate that the conclusions drawn are not influenced by the selectivity.

We construct weights specific to each of the four TwinLife cohorts. First, for a cohort-specific weighting scheme like this, we need to calculate the shares of observations in the TwinLife analysis sample used by highest ISCED and German citizenship of the parents on the household level for each cohort using the categorization presented in Table A1. This share is given by the number of observations in a specific highest ISCED by German citizenship cell (J) for a specific cohort divided by the total number of observations in the analysis sample (N) for a specific cohort. Second, we need to divide the cell-specific correction factors (C) presented in Table A1 by the cohort-specific shares calculated for the analysis sample. The correction factors in Table A1 are based on the cohort-specific shares of observations in the Microcensus multiple-child sample by highest ISCED and German citizenship. Hence, the cohort-specific weights (W) assigned to each observation in the analysis sample depending on highest parental ISCED and parental German citizenship on the household level are given conducting the following calculation:

$$W = C/(J/N) = C \times N/J$$

The resulting weighted analysis sample has the same number of observations as the sample without weights in each cohort but its cohort-specific highest ISCED by German citizenship distribution matches the one in the Microcensus multiple-child sample. If the distributions of parental background indicators for a specific analysis sample based on TwinLife do not differ significantly between the household- and the family-level of aggregation this weighting scheme can also be implemented on the family level.

Table A1 Factors for a selectivity correction weighting scheme based on Microcensus

	Highest educational level (using ISCED) in household			
	1, 2	3a, 3b, 3c	4a, 4b, 5b	5a, 6
Cohort 1				
German citizenship	0.05735661	0.24804655	0.17722361	0.25835412
No German citizenship	0.07547797	0.09293433	0.03142145	0.05918537
Cohort 2				
German citizenship	0.05561700	0.28282509	0.18991942	0.25075051
No German citizenship	0.06794122	0.08547954	0.02938853	0.03807869
Cohort 3				
German citizenship	0.05466035	0.32669826	0.21800948	0.22353871
No German citizenship	0.06082149	0.06650869	0.02085308	0.02890995
Cohort 4				
German citizenship	0.06781795	0.36643281	0.19769743	0.18546501
No German citizenship	0.07609282	0.06817773	0.01978773	0.01852851

Note: The correction factors in the table are not the weights. Please read Appendix A for instructions on how to construct weights using these correction factors.

Sources: Microcensus 2013, own calculations

Information for Authors

Methods, data, analyses (mda) publishes research on all questions important to quantitative methods, with a special emphasis on survey methodology. In spite of this focus we welcome contributions on other methodological aspects.

Manuscripts that have already been published elsewhere or are simultaneously submitted to other journals will not be considered. As a rule we do not restrict authors' rights. All rights remain with the author, and articles in mda are published under the CC-BY open-access license.

Mda aims for a quick peer-review process. All papers submitted to mda will first be screened by the editors for general suitability and then double-blindly reviewed by at least two reviewers. The decision on publication is made by the editors based on the reviews. The editorial team will contact the authors by email with the result at the latest eight weeks after submission; if the reviews have not been received by then, we provide a status update with a new target date.

When preparing a paper for submission, please consider the following guidelines:

- Please submit your manuscript via www.mda.gesis.org.
- The total length of the manuscript shall not exceed 10.000 words.
- Manuscripts should…
  - be written in English, using American English spelling. Please use correct grammar and punctuation. Non-native English speakers should consider a professional language editing prior to publication.
  - be typed in a 12 pt Roman font, double-spaced throughout.
  - be submitted as MS Word documents.
  - start with a cover page containing the title of the paper and contact details / affiliations of the authors, but be anonymized for review otherwise.
  - should be anonymized ("blinded") for review.
- Please also send us an abstract of your paper (approx. 300 words), a front page with a brief biographical note (no longer than 250 words as supplementary file), and a list of 5-7 keywords for your paper.
- Acceptable formats for Graphics are
  - pdf
  - jpg (uncompressed, high quality)
- Please ensure a resolution of at least 300 dpi and take care to send hiqh-quality graphics. Line art images should have a resolution of 500-1000 dpi. Please note that we cannot print color images.
- The type area of our journal is 11.5 cm (width) x 18.5 cm (height). Please consider this when producing tables or graphics.
- Footnotes should be used sparingly.
- Please number the headings of your article. Doing so will make the work of the layout editor easier.

- If your text includes formulas we would like to ask you to upload your text also as a PDF, additional to the Word document.
- By submitting a paper to mda the authors agree to make data and program routines available for purposes of replication.
- Response rates in research papers or research notes, where population surveys are analyzed, should be calculated according to AAPOR standard definitions.
- Please follow the APA guideline when structuring and formating your work.

When preparing in-text references and the list of references please also follow the APA guidelines:

Entire Book:

Groves, R. M., & Couper, M. P. (1998). *Nonresponse in household interview surveys*. New York: John Wiley & Sons.

Journal Article (with DOI):

Klimoski, R., & Palmer, S. (1993). The ADA and the hiring process in organizations. *Consulting Psychology Journal: Practice and Research*, 45(2), 10-36. doi:10.1037/1061-4087.45.2.10

Journal Article (without DOI):

Abraham, K. G., Helms, S., & Presser, S. (2009). How social processes distort measurement: The impact of survey nonresponse on estimates of volunteer work in the United States. *American Journal of Sociology*, 114(4), 1129-1165.

Chapter in an Edited Book:

Dixon, J., & Tucker, C. (2010). Survey nonresponse. In P. V. Marsden & J. D. Wright (Eds.), *Handbook of Survey Research*. Second Edition (pp. 593-630). Bingley: Emerald.

Internet Source (without DOI):

Lewis, O., & Redish, L. (2011). *Native American tribes of Wisconsin*. Retrieved April 19, 2012, from the Native Languages of the Americas website: www.native-languages.org/wisconsin.htm

For more information, please consult the Publication Manual of the American Psychological Association (Sixth ed.).

gesis
Leibniz Institute for the Social Sciences

ISSN 1864-6956 (Print)
ISSN 2190-4936 (Online)