

University of Mannheim

Representativeness and Response Quality of Survey Data

Inaugural dissertation submitted in partial fulfillment of the requirements for
the degree Doctor of Social Sciences in the Graduate School of Economic and
Social Sciences at the University of Mannheim

Submitted by
Carina Cornesse

Mannheim, 10/24/2017

Dekan

Prof. Dr. Michael Diehl

Erstbetreuerin

Prof. Annelies Blom, Ph.D.

Zweitbetreuer

Prof. Joseph Sakshaug, Ph.D.

Erstgutachter

Prof. Joseph Sakshaug, Ph.D.

Zweitgutachter

Prof. Dr. Christof Wolf

Tag der Disputation

24.01.2018

Acknowledgements

There are a number of people who have contributed to this dissertation in different ways and whose help, advice, or general support I would like to acknowledge. First of all, I would like to thank Annelies Blom for her excellent supervision of my thesis and for generally being a great mentor. I am very grateful to her for always providing me with outstanding opportunities and support. I would also like to thank Joe Sakshaug for being a great second dissertation advisor. I am especially grateful for the valuable spot-on feedback that he provided on my early paper drafts. In addition, I would like to thank Christof Wolf for reviewing this thesis and agreeing to chair my dissertation committee. I greatly appreciate the time and effort that my supervisors and reviewers give.

Furthermore, I would like to thank my co-authors: I am grateful to Michael Bosnjak, Tobias Enderle, and Annelies Blom for providing valuable input to our papers. I appreciate and highly value all of my co-authors' contributions to my dissertation. I benefited greatly from their ideas, patience, and expertise.

In general, I would like to express my gratitude to the SFB 884 "Political Economy of Reforms" and GESIS – Leibniz Institute for the Social Sciences for providing an excellent and motivating working environment. I am particularly grateful to my colleagues at the SFB and GESIS. I would especially like to thank Daniela Ackermann-Piek and Suse Helmschrott. I greatly enjoyed working, discussing, laughing, and sometimes moaning with them. I would also like to thank Christian Bruch, Barbara Felderer, Franziska Gebhard, Jessica Herzing, Marina Jesse, Uli Krieger, Rita Maklakova, and Dayana Sieger. I highly appreciate the valuable and inspiring discussions we had at work and I am grateful for their moral support. At GESIS, I would like to thank all former and current members of the GESIS Panel team.

Finally, I am grateful to my family, especially my boyfriend Sebastian Marx and my parents Alfred and Elvira Cornesse for their encouragement, support, and patience.

Table of contents

- General introduction 9
 - 1. The relevance of survey data quality 9
 - 2. The Total Survey Error (TSE) framework 11
 - 3. The concept of survey representativeness 13
 - 4. The concept of survey response quality..... 15
 - 5. Summary of dissertation papers 16
- Literature 21
- Is there an association between survey characteristics and representativeness? A meta-analysis..... 27
 - Abstract 27
 - Keywords 27
 - Acknowledgements 27
 - 1. Background and aims 29
 - 2. Survey representativeness concepts..... 30
 - 3. Conceptual development of research question and expectations 32
 - 3.1 Probability surveys versus nonprobability surveys 32
 - 3.2 Response rates 34
 - 3.3 Mixed-mode surveys versus single-mode surveys 35
 - 3.4 Web surveys versus other single-mode surveys 36
 - 3.5 Auxiliary variables..... 37
 - 4. Method 38
 - 4.1 Literature search, study identification and data extraction..... 38
 - 4.2 Meta-analytic procedure 39
 - 4.3 Effect size measures 40
 - 5. Results 42
 - 5.1 General findings..... 43
 - 5.2 Moderator analyses..... 43
 - 6. Summary and conclusion 49
 - Literature 52
 - Online Appendix A: R-Indicator articles in the meta-analysis 60
 - Online Appendix B: Benchmark comparison articles in the meta-analysis..... 63
- Is it in the method? Testing five measures of survey representativeness..... 71
 - Abstract 71
 - Keywords 71

Acknowledgements	72
1. Introduction.....	73
2. Measures of survey representativeness	75
2.1 Representativeness measures at the aggregate level.....	79
2.2 Representativeness measures at the variable level.....	83
2.3 Representativeness measures at the category level.....	85
3. Data	86
3.1 The survey designs of the GIP 2012, the GESIS Panel, and the GIP 2014: differences and similarities	86
3.2 Auxiliary data for modelling representativeness	89
4. Results	92
4.1 Response rates	93
4.2 R-Indicators	94
4.3 Fraction of Missing Information (FMI)	95
4.4 Subgroup response rates.....	97
4.5 Benchmark comparison.....	101
5. Discussion	103
Literature	109
Online Appendix	116
The utility of auxiliary data for survey response modeling: Evidence from the German Internet Panel	125
Abstract	125
Keywords	125
Acknowledgements	125
1. Introduction.....	126
2. Uses and usefulness of auxiliary data	127
3. Types of auxiliary data and their quality	129
4. Data and methods	132
4.1 The German Internet Panel (GIP)	132
4.2 The auxiliary data	133
4.3 Methods	135
5. Results	136
5.1 Missing data	136
5.2 Correlations with survey response.....	137
5.3 Survey response models.....	138

6. Discussion	141
Literature	144
Response quality in nonprobability and probability-based online panels	151
Abstract	151
Keywords	151
Acknowledgements	151
1. Introduction.....	153
2. Respondent motivation and survey satisficing	154
3. Measuring satisficing.....	156
3.1 Straight-lining	158
3.2 Item nonresponse.....	158
3.3 Midpoint selection.....	159
4. Online panel data	160
4.1 The GIP	160
4.2 The GP	161
4.3 Panel 1	162
4.4 The nonprobability online panels: Panel 2 - Panel 8.....	162
5. Methods	163
5.1 Straight-lining	163
5.2 Item nonresponse.....	163
5.3 Midpoint selection.....	165
6. Results	167
6.1 Straight-lining	167
6.2 Item nonresponse.....	168
6.3 Midpoint selection.....	171
6.4 Response quality and costs	173
7. Discussion	173
Literature	177
Appendix A: Panel costs	183
Appendix B: Question texts and answer scales by indicator.....	184
Appendix C: Midpoint design experiment by panel	195
General conclusion	199
General Appendix: Eidesstattliche Erklärung	204

General introduction

General introduction

1. The relevance of survey data quality

In the social sciences, research is often based on findings from survey data. Common research topics examine political behavior, societal attitudes and opinions, as well as personal values. Survey results shape societal debates and can have an impact on policy decisions. A recent example for the impact of survey results is the debate about same-sex marriage in Germany. A report by the German Federal Anti-Discrimination Agency released in early 2017 reported that 82.4% of the German population was in favor of allowing same-sex marriage based on a large-scale telephone survey (see Antidiskriminierungsstelle des Bundes, 2017). When in June 2017 a short but heated discussion about passing a law that would allow same-sex marriage took place in the German public sphere and the parliament, this survey finding was a prominent argument in favor of same-sex marriage in the news coverage¹. Considered as a valid expression of the German public view on this topic, this survey result might have helped proponents of the admission of same-sex marriage to pass the respective law in the German parliament by the end of June 2017 in a sudden and uncharacteristically fast legislative process.

However impactful the results might be, research and policy debates based on survey data rely on the assumption that the survey data are of high enough quality to be able to draw inferences from the data to a broader population. However, collecting high quality survey data is challenging. Common methodological issues include declining response rates (see e.g., de Heer & de Leeuw, 2002) as well as concerns about biases due to systematic misrepresentation of members of the target population (see e.g. Groves & Lyberg, 2010) as well as measurement error (see e.g., O'Muirheartaigh, 1997).

That survey data can be wrong has recently been shown repeatedly in the area of election polling.

Prominent in British news coverage, for example, were the mispredictions of many polls with regard

¹ See for example <http://www.handelsblatt.com/politik/deutschland/kehrtwende-der-kanzlerin-merkel-macht-ehe-fuer-alle-ploetzlich-moeglich/19983854.html>.

to the 2015 general election². Most polls had predicted the Conservative Party to be tied with Labour. Yet the final election outcome was a clear win for the Conservatives. Similarly, most British polls predicted that the British public would vote to remain in the European Union in the 2016 referendum. Yet the outcome of the referendum was that Britain would leave the EU³. Outside of Britain, too, media coverage of election poll mispredictions has occurred, for example in the United States where many polls predicted Hillary Clinton to win the presidential election in 2016 while the final outcome was that Donald J. Trump won the race⁴. Other examples of failed predictions from survey data include failure to predict voter turnout (see for example Holbrook & Krosnick, 2010), income (see for example Hariri & Lassen, 2017), and religious attendance (see for example Presser & Stinson, 1998).

Because results from survey data can be inaccurate and possibly lead to wrong predictions, it is important to ask whether survey data have the necessary quality to be able to draw valid inferences. This question is, however, often difficult to answer because multiple error sources can influence survey data quality. Among other factors, the quality of the survey data is influenced by survey design characteristics, such as the sampling method and the survey mode. In an inquiry report, Sturgis et al. (2016), for instance, find that the main reason for the failure of most election polls in the 2015 UK general election was “unrepresentative samples” (p. 69; for similar findings on mispredictions of election polls see AAPOR, 2009).

It is imperative to ensure that research findings from survey data are valid by avoiding the different types of survey errors that can arise throughout the survey process. In order to do this, researchers need to keep the various error sources in mind and be able to detect them. Such an overview of potential survey errors can be gained by applying the Total Survey Error framework.

² See for example <http://www.bbc.com/news/uk-politics-32751993>.

³ See for example <https://www.cnbc.com/2016/07/04/why-the-majority-of-brexit-polls-were-wrong.html>.

⁴ See for example <https://www.nytimes.com/2017/05/31/upshot/a-2016-review-why-key-state-polls-were-wrong-about-trump.html>.

2. The Total Survey Error (TSE) framework

The TSE provides a framework to examine the various errors that might occur in the survey process (Groves, 2004; Groves & Lyberg, 2010). As components of the survey process, the TSE includes measurement steps as well as representational inference steps. The measurement steps of the TSE (see left side of Figure 1) start with the operationalization of theoretical constructs into survey questions as measurements. The next step of the measurement process is the retrieval of responses to the survey questions. Then, the responses are processed into edited data. The representational inference steps of the TSE (see right side of Figure 1) start with the identification of the target population of the survey from the inferential population. The next step is the construction of the sampling frame that is supposed to contain all members of the target population. Then, a sample is drawn from the sampling frame. After this, a set of respondents is generated from the sample by conducting survey interviews. The end result of the measurement and representational inference steps of the TSE is the survey statistic, which is calculated from the edited data on the set of respondents.

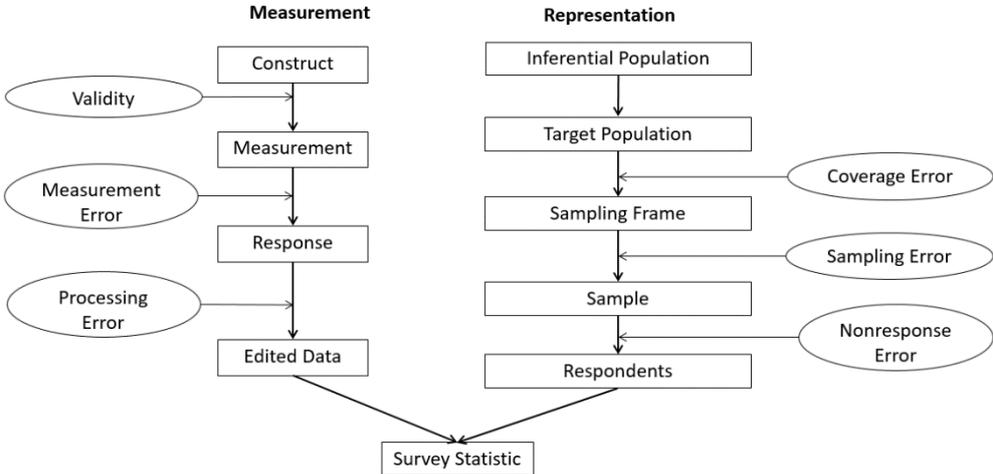


Figure 1: Total Survey Error Components Linked to Steps in the Measurement and Representational Inference Process (taken from Groves et al., 2004).

At each step of the TSE process, errors can occur. Regarding the measurement steps of the TSE framework, the construct might be erroneously operationalized into a survey question. This might result in invalid measurements. In addition, measurement error can arise in the measurement

process resulting in incorrectly measured items that do not reflect the true population values. Furthermore, processing error might occur in the process of recording and coding the survey responses into the edited data set.

In the representational inference steps of the TSE framework, coverage error can occur when the sampling frame either excludes members of the target population or includes nonmembers of the target population. This can lead to biased estimates if there is systematic under-coverage or over-coverage of certain parts of the target population. In addition, sampling error can arise because not every member of the sampling frame is included in the sample. This often leads to imprecisions in the survey estimates. Furthermore, nonresponse error can occur when members of the sample do not respond to the survey. As an additional error source in the survey process, researchers often discuss adjustment error, which arises when statistical adjustments for misrepresentations in the data fail to correct these misrepresentations.

All errors that can arise throughout the Measurement and Representational Inference Process of the TSE framework described above can result in incorrect or imprecise survey statistics. Therefore, it is necessary to examine survey data from different perspectives keeping in mind that both the measurement steps and the representational steps of the TSE framework influence the survey data. In the first three dissertation papers, I examine the representational inference steps of the TSE from different angles by applying different types of definitions of the concept of survey representativeness. In my last dissertation paper, I turn the focus to the measurement steps of the TSE framework by applying the concept of response quality as a potential type of measurement error.

3. The concept of survey representativeness

The concept of survey representativeness refers to the representational inference steps of the TSE framework. It generally includes coverage error, sampling error, nonresponse error, and adjustment error as potential error sources. There is not one agreed-upon definition of survey representativeness. Instead, the term is often used ambiguously (see Kruskal & Mosteller, 1979a, 1979b, 1979c for a collection of definitions) and seldom defined explicitly.⁵ Because it is used ambiguously, some researchers resent using the term “representativeness” altogether (see for example Rendtel & Pötter, 1992). A common alternative term is “survey accuracy” (see for example Yeager et al., 2011).

In the current scientific literature on survey representativeness, two definitions of survey representativeness are commonly used, even if mostly only implicitly. The first definition refers to the similarity between the response set of a survey compared to the target population (see left side of Figure 2 for an illustration). Although this is an intuitive definition, it is hard to operationalize it in a measurement. Usually, studies working with this underlying definition compare a survey response set to a benchmark that is assumed to accurately reflect the target population. Such benchmarks can be official statistics such as a census or large-scale surveys that are assumed to be of high quality.

An advantage of this approach is that no knowledge about the nonrespondents is necessary to conduct this type of representativeness assessment. Drawbacks include that the assumption of an accurate benchmark might be violated and that in many cases only socio-demographic characteristics can be compared because these are the only available benchmark data.

This type of studies can commonly be found in fieldwork reports (see for example European Social Survey, 2016) and methodological literature on survey mode effects (see for example Malhotra & Krosnick, 2007) as well as post-survey adjustment weights (see for example Steinmetz et al., 2014).

⁵ A rare exception is the paper by Schouten et al. (2009), where representativeness is defined as a “respondent selection” that is “as close as possible to a ‘simple random sample’” (p. 103). In the current literature, the term “representativeness” is often used to say that a survey is of high general quality.

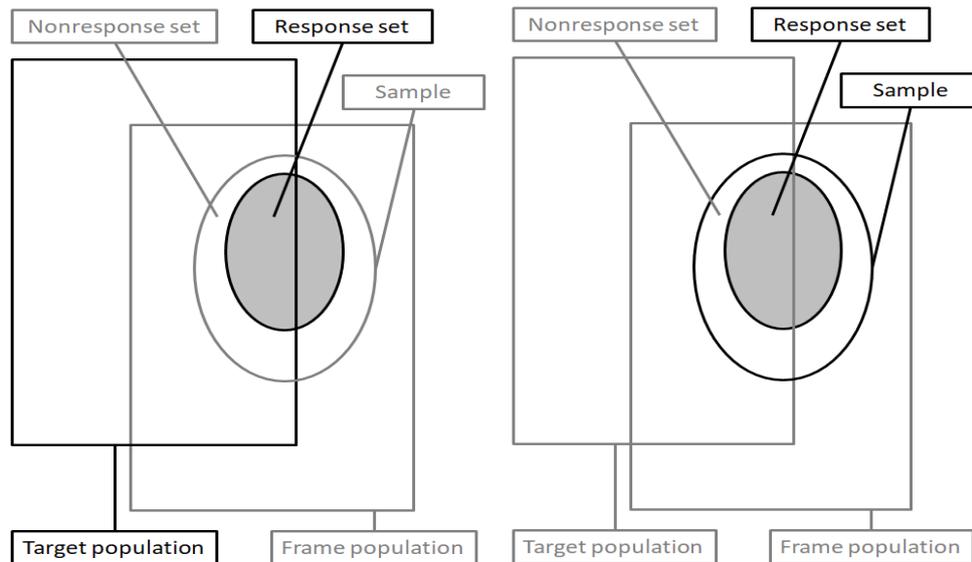


Figure 2: Strict and weak definition of survey representativeness (adapted from Särndal and Lundström, 2005)

The second definition of survey representativeness that is used in large parts of the scientific literature relies on the assumption that the gross sample of a probability-based survey correctly represents the target population (see right side of Figure 2 for an illustration). Therefore, representativeness can be assessed by comparing the response set to the gross sample of the survey, which includes the survey respondents as well as the nonrespondents. Strictly speaking, this concept is equal to that of nonresponse bias.

An advantage of this approach is that in many cases a variety of variables are either available for the representativeness assessment directly, e.g. from the sampling frame of the survey, or additional data can be linked to the available data from other sources, such as commercial or official area data. A potential drawback is that it is necessary to obtain data for the survey nonrespondents, which might be challenging if data-rich sampling frames are not available. In addition, indicators that rely on this representativeness approach neglect potential sampling biases, for example due to sampling frame imperfections.

This type of studies can commonly be found in the statistical literature on model-based data quality measures (see for example Schouten et al., 2009). They can also be found in the literature on

fieldwork monitoring (see for example Schouten, Shlomo, & Skinner, 2011) and tailored fieldwork designs (see for example Luiten & Schouten, 2013).

Regardless of the definition, survey representativeness is necessary for obtaining accurate survey statistics following the TSE framework. Therefore, research on the measurement of survey representativeness and the survey characteristics associated with it is crucial. This dissertation contributes to this line of research.

4. The concept of survey response quality

The concept of survey response quality refers to the idea that respondents might provide better or worse responses depending on the quality of the respondents' response processing. The dominant response processing model in the current literature was developed by Tourangeau, Rips, and Rasinski (2000). It describes the process from the question stimulus to the survey response. The cognitive steps involved in response processing are: question comprehension, information retrieval, judgment and estimation, and reporting an answer (see Figure 3).

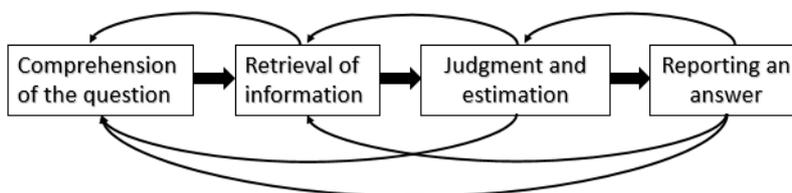


Figure 3: A simple model of the survey response process (adapted from Groves et al., 2009).

In the question comprehension step, respondents interpret the question. They process the words and structure of the question, assign meaning to the question, and evaluate what information the question asks for. The retrieval of information step is where respondents recall the necessary information they need to answer the question. The information retrieval is set off by cues in the question that trigger recall from memory. The respondent's memory then provides further cues that help in answering the question. This process might go back and forth between question cues and memory cues several times until the necessary information is retrieved. In the judgment and estimation step of the response process, the respondent combines and summarizes the recalled

information. When reporting an answer, the respondent formulates the response in the required format. This involves mapping the actual answer to the response options. While there is a certain plausibility to the order of the cognitive processes, starting with question comprehension, and ending with the reporting of an answer, the processes are not always carried out in that order. Instead, there can be a substantial amount of backtracking from one step to another, as well as an overlap between cognitive steps.

Following the prominent satisficing theory developed by Krosnick (1991), respondents at each of the processing steps obtain either optimized results that are as best as possible or satisficed results that only seem satisfactory. If respondents satisfice, this means that they take short-cuts through the optimal cognitive response process to reduce the cognitive burden (Krosnick & Alwin, 1987). They might for instance only skim the question text for keywords, they might skip reading definitions and other supplementary material meant to help them answer the question correctly, or they might only browse through the answer options until they find an option that fits them reasonably well without checking whether there is another answer option that fits them even better.

If respondents satisfice rather than optimize, response quality is compromised and measurement error is introduced into the data. Following the TSE framework, this can result in erroneous survey statistics. Therefore, research on the measurement of response quality and the survey characteristics associated with it is crucial. This dissertation contributes to this line of research.

5. Summary of dissertation papers

My dissertation consists of four papers. The first paper is a meta-analysis on the association between survey characteristics and representativeness in the published literature. The second paper is an examination of the informative value of five common measures of survey representativeness using the recruitment data of two large-scale probability-based online panels in Germany. The third paper is an assessment of the utility of auxiliary data for predicting survey response in a probability-based online panel. The last paper explores the effects of respondent motivation on response quality in

three probability-based and seven nonprobability online panels in Germany. With these four papers, my dissertation covers potential errors in survey data discussed in the TSE framework from different perspectives.

Paper 1: Is there an association between survey characteristics and representativeness? A meta-analysis

In this paper, I synthesize the literature on survey representativeness and its association with common survey design characteristics. The paper gives an overview of common definitions of the term representativeness and sheds light on ongoing debates about how it is related to the design of a survey. In particular, the paper discusses the association of survey representativeness with the sampling design, the survey mode, the response rate, and the auxiliary data used in the representativeness assessment.

In addition to providing an overview of the state of the literature with regard to these methodological issues, I present empirical evidence on how these survey design characteristics are related to representativeness. In this empirical assessment, I apply meta-analytic techniques to synthesize the results that we find in the published literature on two common measures of survey representativeness: R-Indicators and descriptive benchmark comparisons.

The results from the meta-analytic moderator analyses indicate that probability-based samples, mixed-mode surveys, and other-than-Web mode surveys are more representative than nonprobability samples, single-mode surveys, and web surveys. In addition, I find that there is a positive association between representativeness and the response rate.

A key finding of this paper is, however, that more research into the topic of survey representativeness is needed. In the literature, there is, for instance, a strong need for studies that assess survey representativeness on more than one measure. There are also only few studies that assess the representativeness of online surveys using benchmark comparisons and no studies that assess representativeness of web surveys using R-Indicators. In general, I find that more primary

research is needed to enhance the informative value of meta-analyses in order to derive recommendations for survey practitioners aiming to design a generally representative survey.

Paper 2: Is it in the method? Testing five measures of survey representativeness

The second paper assesses and discusses the informative value of five common measures of survey representativeness for use in a cross-survey as well as within-survey representativeness comparison. I test measures on different levels of aggregation. On the most aggregated level, I test response rates and R-Indicators as single-number representativeness overviews. On the variable level, I assess the value of Fractions of Missing Information (FMIs) to provide evidence on the representativeness of survey data. On the category level, I examine subgroup response rates as well as benchmark comparisons and assess their added value to the representativeness analysis.

The data I use in this study are from two probability-based online panels in Germany: the German Internet Panel (GIP) and the GESIS Panel. Both panels share a number of survey design characteristics but also differ in other aspects. I assess the informative value of each representativeness measure for a comparison across the panels. In addition, I compare the representativeness measures within each panel across its recruitment stages.

My findings show that to get a broader picture of survey representativeness, it is necessary to apply multiple measures at different levels of aggregation. When identical auxiliary data are available on the surveys that researchers want to compare, the R-Indicators provide a more meaningful representativeness assessment than the response rates because they take more information into account. FMIs are difficult to apply in a comparative context because they overemphasize the relative size of the response set compared to the overall gross sample in a way that distorts survey comparisons. Both subgroup response rates and benchmark comparisons add value to the representativeness assessments by providing more detail than the aggregate measures. When only using these category-level measures, however, it is hard to make a general representativeness

assessment. Therefore, I conclude that using a variety of measures on different levels of aggregation is advisable to provide a comprehensive picture of the broad concept of survey representativeness.

Paper 3: The utility of auxiliary data for survey response modeling: Evidence from the German Internet Panel

In this paper, I explore the value of different sources of auxiliary data (sampling frame data, interviewer observations, and commercial as well as official micro-geographic area data) in the context of the recruitment of the GIP, a probability-based online panel in Germany. I examine the proportions of missing values in the auxiliary variables available in this study and investigate whether there are systematic differences in the proportions of missing values between respondents and nonrespondents. In addition, I assess the correlations of the auxiliary variables with survey response as well as the predictive power and significance of coefficients in survey response models. I conduct all of the analyses on survey response in the GIP face-to-face recruitment interview as well as the subsequent online profile survey.

I find in this study that all of the auxiliary data have problems: the sampling frame data are scarce, the interviewer observations have relatively high proportions of missing values that are in part systematically missing by survey response, the Microm data are intransparent and have some missing data as well, and the INKAR data are highly aggregated. In addition, none of the auxiliary variables correlate to any substantial degree with survey response at the two GIP recruitment steps. Therefore, none of the auxiliary data have predictive power in the survey response models although some of the coefficients in the models are significant.

I conclude that the auxiliary data examined in this study should be used with caution in survey practice. In addition, more research is needed into the quality of different sources of auxiliary data. Furthermore, this study shows that the search for high quality auxiliary data that are predictive of survey response should continue to fill the need for these data in survey operations and methodological research.

Paper 4: Response quality in nonprobability and probability-based online panels

The last paper of my dissertation discusses the topic of response quality across nonprobability and probability-based online panels. The basic assumption in this paper is that monetary rewards are a much more important motivation for participation in nonprobability online panels than in probability-based online panels. This is because nonprobability online panels recruit their respondents primarily by online advertisements that promise fast cash in return for filling out questionnaires. Due to the self-selection mechanism that nonprobability online panels rely on for panel recruitment, it is likely that this strategy has an impact on the motivational composition of the set of people that self-selects into these nonprobability panels. While monetary rewards likely motivate probability-based online panel participants as well, I assume that there is a wider variety of motivations for participation in probability-based online panels.

Based on these assumptions, I compare the response quality of seven nonprobability online panels to the response quality of three probability-based online panels. All online panels fielded the same short survey at approximately the same fieldwork period. The response quality indicators I examine in the data of the ten online panels are straight-lining, item nonresponse, and midpoint selection in a visual design experiment. The results show that there is significantly more straight-lining in the nonprobability online panels than in the probability-based online panels. However, I find no systematic pattern indicating that response quality is lower in nonprobability online panels than in probability-based online panels with regard to item nonresponse and the midpoint selection experiment. I conclude that more research is needed into response quality in nonprobability online panels.

Literature

AAPOR (2009). *An Evaluation of the Methodology of the 2008 Pre-Election Primary Polls*. American Association for Public Opinion Research.

Antidiskriminierungsstelle des Bundes (2017). *Einstellungen gegenüber Lesben, Schwulen und Bisexuellen in Deutschland. Ergebnisse einer bevölkerungsrepräsentativen Umfrage*. Retrieved from http://www.antidiskriminierungsstelle.de/SharedDocs/Downloads/DE/publikationen/Umfragen/Handout_Themenjahrumfrage_2017.pdf?__blob=publicationFile&v=3.

Curtin, R., Presser, S., & Singer, E. (2005). Changes in telephone survey nonresponse over the past quarter century. *Public Opinion Quarterly*, 69(1), 87-98.

De Heer, W., & De Leeuw, E. (2002). Trends in household survey nonresponse: A longitudinal and international comparison. In Groves, R. M., Dillman, D. A., Eltinge, J. L., Little, R. J. A. (Eds.), *Survey nonresponse* (pp.41-54). New York: Wiley.

DeMaio, T. J. (1984). Social desirability and survey. In Turner, C. F. & Martin, E. (Series Eds.), *Surveying subjective phenomena: Vol. 2*. (pp. 257-282). Russel Sage Foundation.

Eisenhower, D., Mathiowetz, N. A. & Morganstein, D. (1991). Recall Error: Sources and Bias Reduction Techniques. In Biemer, P. P., Groves, R. M., Lyberg, L. E., Mathiowetz, N. A., Sudman, S. (Eds.), *Measurement Errors in Surveys* (pp. 127-145). John Wiley & Sons, Inc.

European Social Survey (2016). *ESS7-2014 Documentation Report. Edition 3.0*. Bergen: European Social Survey Data Archive, Norwegian Social Science Data Services for ESS ERIC. Retrieved from http://www.europeansocialsurvey.org/docs/round7/survey/ESS7_data_documentation_report_e03_0.pdf.

Groves, R. M., Fowler, F. J., Couper, M. P., Lepkowski, J. M., Singer, E., Tourangeau, R. (2009). *Survey methodology*. John Wiley & Sons, Inc.

- Groves, R. M., & Lyberg, L. (2010). Total Survey Error: Past, Present, and Future. *Public Opinion Quarterly*, 74(5), 849-879.
- Groves, R. M., Presser, S., Tourangeau, R., West. B. T., Couper, M. P., Singer, E., Toppe, C. (2012). Support for the Survey Sponsor and Nonresponse Bias. *Public Opinion Quarterly* 76(3), 512-524.
- Hariri, J. G., & Lassen, D. D. (2017). Income and Outcomes: Social Desirability Bias Distorts Measurements of the Relationship between Income and Political Behavior. *Public Opinion Quarterly*, 81(2), 564-576.
- Holbrook, A.L., & Krosnick, J.A. (2010). Social desirability bias in voter turnout reports: tests using the item count technique. *Public Opinion Quarterly*, 74, 37–67.
- Krosnick, J. A. (1991). Response Strategies for Coping with the Cognitive Demands of Attitude Measures in Surveys. *Applied Cognitive Psychology*, 5(3), 213–236.
- Krosnick, J. A., & Alwin, D. F. (1987). An evaluation of a cognitive theory of response-order effects in survey measurement. *Public Opinion Quarterly*, 51(2), 201–219.
- Kruskal, W., & Mosteller, F. (1979a). Representative sampling, I: Non-scientific literature. *International Statistical Review/Revue Internationale de Statistique*, 47(1), 13-24.
- Kruskal, W., & Mosteller, F. (1979b). Representative sampling, II: Scientific literature, excluding statistics. *International Statistical Review/Revue Internationale de Statistique*, 47(2), 111-127.
- Kruskal, W., & Mosteller, F. (1979c). Representative sampling, III: The current statistical literature. *International Statistical Review/Revue Internationale de Statistique*, 47(3), 245-265.
- Luiten, A., & Schouten, B. (2013). Tailored fieldwork design to increase representative household survey response: an experiment in the Survey of Consumer Satisfaction. *Journal of the Royal Statistical Society*, 176(1), 169–189.

Malhotra, N., & Krosnick, J. A. (2007). The Effect of Survey Mode and Sampling on Inferences about Political Attitudes and Behavior: Comparing the 2000 and 2004 ANES to Internet Surveys with Nonprobability Samples. *Political Analysis*, 15(3), 286–323.

O'Muircheartaigh, C. (1997). Measurement Error in Surveys: A Historical Perspective. In Lyberg, Biemer, Collins, de Leeuw, Dippo, Schwarz, Trewin (Eds.). *Survey Measurement and Process Quality*. John Wiley & Sons, Inc.

Presser, S., & Stinson, L. (1998). Data Collection Mode and Social Desirability Bias in Self-Reported Religious Attendance. *American Sociological Review*, 63(1), 137-145.

Rendtel, U., & Pötter, U. (1992). *Über Sinn und Unsinn von Repräsentationsstudien* (No. 61). DIW Discussion Papers.

Schouten, B., Cobben, F., & Bethlehem, J. (2009). Indicators for the representativeness of survey response. *Survey Methodology*, 35(1), 101-113.

Schouten, B., Shlomo, N., & Skinner, C. (2011). Indicators for Monitoring and Improving Representativeness of Response. *Journal of Official Statistics*, 27(2), 1-24.

Steinmetz, S., Bianchi, A., Tijdens, K., & Biffignandi, S. (2014). Improving Web Survey Quality: Potentials and Constraints of Propensity Score Adjustments. In Callegaro, M., Baker, R., Bethlehem, J., Göritz, A. S., Krosnick, J. A., and Lavrakas, P. J. (Eds.), *Online Panel Research. A Data Quality Perspective* (pp. 273–298). Chichester, UK: John Wiley & Sons.

Sturgis, P., Baker, N., Callegaro, M., Fisher, S., Green, J., Jennings, W., Kuha, J., Lauderdale, B., & Smith, P. (2016). *Report of the Inquiry into the 2015 British general election opinion polls*. London: Market Research Society and British Polling Council.

Tourangeau, R., Rips, L. J., & Rasinski, K. A. (2000). *The psychology of survey response*. Cambridge, New York: Cambridge University Press.

Yeager, D. S., Krosnick, J. A., Chang, L., Javitz, H. S., Levendusky, M. S., Simpser, A., & Wang, R. (2011). Comparing the accuracy of RDD telephone surveys and internet surveys conducted with probability and non-probability samples. *Public Opinion Quarterly*, 75(4), 709-747.

Paper 1

Is there an association between survey characteristics and representativeness? A meta-analysis

Paper 1

Is there an association between survey characteristics and representativeness? A meta-analysis⁶

Abstract

How to achieve survey representativeness is a controversially debated issue in the field of survey methodology. Common questions include whether probability-based samples produce more representative data than nonprobability samples, whether the response rate determines the overall degree of survey representativeness, and which survey modes are effective in generating highly representative data. This meta-analysis contributes to this debate by synthesizing and analyzing the literature on two common measures of survey representativeness (R-Indicators and descriptive benchmark comparisons). Our findings indicate that probability-based samples (compared to nonprobability samples), mixed-mode surveys (compared to single-mode surveys), and other-than-Web modes (compared to Web surveys) are more representative, respectively. In addition, we find that there is a positive association between representativeness and the response rate. Furthermore, we identify significant gaps in the research literature that we hope might encourage further research in this area.

Keywords

Meta-analysis, survey representativeness, R-indicator, descriptive benchmark comparisons, response rate, nonprobability sampling, mixed mode, web surveys, auxiliary data

Acknowledgements

The authors gratefully acknowledge support from the Collaborative Research Center (SFB) 884 “Political Economy of Reforms” (project A8), funded by the German Research Foundation (DFG) and

⁶ This paper is joint work with Michael Bosnjak. The current version has been accepted for publication in Survey Research Methods.

from GESIS – Leibniz Institute for the Social Sciences. The authors would like to especially thank Barry Schouten, Joe Sakshaug, and the participants of the International Workshop on Household Survey Nonresponse 2016 in Oslo, Norway, for their feedback on early versions of this paper. We would also like to thank Christian Bruch and Barbara Felderer for their valuable feedback on variance estimation and our student assistants Sophia Fauser, Margarita Kozlova, Linda Beck, and Thomas Alcock for their help with coding the data.

1. Background and aims

One of the most important questions in the research field of survey methodology is how to collect high quality survey data that can be used to draw inferences to a broader population. Extensive research has been conducted from different angles, often creating an ambiguous picture of whether and how a specific survey characteristic might help or hurt in the pursuit of reaching high survey quality. An example is the current debate around the representativeness of probability-based surveys versus nonprobability surveys, where proponents and opponents of each approach provide new findings on a regular basis. Some of these studies suggest that probability-based surveys are more accurate than nonprobability surveys (e.g., Malhotra & Krosnick, 2007; Loosveldt & Sonck, 2008; Chang & Krosnick 2009; Yeager et al., 2011). Other studies demonstrate that nonprobability surveys are as accurate as or even more accurate than the probability-based surveys (e.g., Gelman et al., 2017; Wang et al., 2015).

Another example of a scientific discussion in the survey methodological community is the question of whether the response rate can be used as a representativeness indicator. Even though meta-analytic research shows that the response rate is only weakly associated with nonresponse bias (Groves & Peytcheva, 2008; Groves et al., 2008) the debate about how to stop the decrease in response rates goes on (e.g., de Leeuw & de Heer, 2002; Brick & Williams, 2013). Another topic continuously investigated in survey methodological research with seemingly contradictory findings is which survey mode and which mode combinations yield the most representative data (e.g., Luiten & Schouten, 2012). Based on a systematic review summarizing the available meta-analytic evidence about mode effects and mixed-mode effects on representativeness, Bosnjak (2017, p. 22) concludes: “there is a lack of meta-analytic studies on the impact of (mixing) modes on estimates of biases in terms of measurement and representation, not just on proxy variables or partial elements of bias, such as response rates.”

In light of the extensive research on different aspects of survey representativeness that partly leads to seemingly contradictory findings, the overall aims of this paper are to identify and systematically

synthesize the existing literature on survey representativeness, and to answer the question if, and to what extent, specific survey characteristics are associated with survey representativeness. Before we develop the specific research question and hypotheses of this study, we define the scope by specifying the two survey representativeness concepts considered.

2. Survey representativeness concepts

Survey representativeness is an ambiguous and controversial term. In survey methodology, it most commonly refers to the success of survey estimates to mirror ‘true’ parameters of a target population (see Kruskal & Mosteller, 1979a, 1979b, 1979c for this and other definitions of the concept of representativeness). The fact that a survey can be representative regarding the variables of interest to one researcher, and misrepresent the variables of interest to another researcher at the same time, adds to the vagueness of the concept of a generally representative survey. Since some researchers reject the term ‘representativeness’ entirely (e.g., Rendtel & Pötter, 1992; Schnell, 1993), research on the subject frequently emerges under the keywords ‘accuracy’ or ‘data quality’ in general (e.g. Malhotra & Krosnick, 2007; Yeager et al., 2011). Other researchers still use the term ‘representativeness’ (e.g., Chang & Krosnick, 2009) or ‘representativity’(e.g., Bethlehem et al., 2011). From a total survey error perspective, survey non-representativeness encompasses the errors of nonobservation: coverage error, sampling error, nonresponse error, and adjustment error (Groves et al., 2004; Groves & Lyberg, 2010).

Because of its ambiguity, there are various operationalizations of survey representativeness. Two common approaches prevail; each is associated with specific advantages and disadvantages. The first approach is to compare the survey response set to the target population in question. This is mostly done by using benchmark comparisons, estimating the absolute bias. For this type of representativeness measure, it is necessary to assume that the benchmark to which the response set is compared is unbiased and reflects the ‘true’ values of the target population. Therefore, these representativeness assessments are mostly conducted on socio-demographic characteristics only (e.g., Keeter et al., 2000; Wozniak, 2016). The goal is in many cases to show whether a specific data

source is trustworthy and can therefore be used to answer a substantive research question, for example on health (Potthof et al., 2004) road safety (Goldenbeld & de Craen, 2013) or religion and ethnicity (Emerson et al., 2010).

The other approach to operationalizing the concept of representativeness is to compare the survey response set to the gross sample including respondents as well as nonrespondents. This allows the inclusion of a much broader set of variables in the assessment compared to the absolute bias measure. For instance, data from the sampling frame, fieldwork, and from data linkage can be used for the representativeness assessment. Examples of such measures include R-Indicators (Schouten et al., 2009), balance and distance measures (Lundquist & Särndal, 2013), and Fractions of Missing Information (Wagner, 2010). However, since only those characteristics can be assessed that are available for respondents as well as for nonrespondents, data availability is an issue also for measures operationalizing this approach. In addition, the underlying assumption is that if the response set perfectly reflects the gross sample, including both respondents and nonrespondents, this constitutes a representative survey. This assumption leaves out the fact that the sample might already be biased compared to the target population due to coverage error and sampling error. Therefore, these measures operationalizing the second approach to representativeness are – technically speaking – nonresponse bias measures (Wagner, 2012).

Because there is no agreed-upon measure of survey representativeness, we focus on two conceptualizations that are common in the survey methodological literature. First, we investigate R-Indicators that assess representativeness by comparing respondents to the gross sample of a survey, which includes respondents as well as nonrespondents. Second, we examine descriptive benchmark comparisons that measure survey representativeness by comparing respondent characteristics to an external benchmark that is supposed to reflect the characteristics of the target population.

3. Conceptual development of research question and expectations

In this paper, we assess whether there is an association between survey characteristics and the reported degree of representativeness of a survey. We focus on a selection of survey characteristics that are commonly considered to affect survey representativeness in the current research literature and for which we could gather sufficient information during the data extraction process. Specifically, we expect that the following five survey characteristics are associated with the reported degree of representativeness: probability surveys versus nonprobability surveys, response rates, mixed-mode surveys versus single-mode surveys, web surveys versus offline surveys, and the number of auxiliary variables (i.e., data that is available for respondents as well as nonrespondents) used in the representativeness assessments. In the following, we describe our expectations for each survey characteristic considered in more detail. These associations should hold regardless of whether representativeness is operationalized using R-Indicators (an indicator based on response propensity models using auxiliary data; see section 4.3.1 in this paper for more details) or descriptive benchmark comparisons.

3.1 Probability surveys versus nonprobability surveys

Probability sampling theory states that valid inference and the assessment of an estimate's uncertainty can only be guaranteed by random selection of units of analysis from an accurate sampling frame (e.g. Kish, 1965; Lohr, 2010). Some researchers have, however, suggested that true probability sample surveys of the population do not exist in practice due to non-random survey nonresponse (e.g., Gelman et al., 2017).

Over the last decades, the number of nonprobability surveys, especially commercial opt-in online panels, has increased immensely. Their advantages include that they are fast and relatively cheap. Furthermore, they often reach high numbers of respondents (e.g., Bethlehem & Biffignandi, 2012). They also often have respondent pools containing millions of people, sometimes even across several

countries, as for instance Google 360⁷, Research Now⁸, Omnirussia⁹, or Toluna¹⁰. In many cases, little information is available on how these nonprobability panels recruit people, how many people are in their respondent pool, how they sample from the respondent pool to conduct an individual survey, and how many people in their respondent pool are actually active, i.e. respond to survey requests regularly. Furthermore, it has been observed that active panelists often participate in multiple non-probability panels (e.g. Vonk et al., 2006; Tourangeau et al., 2013), especially since there are databases that contain many opt-in panels from which people who would like to earn money by completing surveys can choose as many panels as they like (e.g., www.umfragenplatz.de; www.mysurvey.com/). In addition, a Google search performed in July 2017 for the term “money for surveys” yielded 180.000 hits and the application Google Opinion Rewards, that offers Google Play credit in return for answering surveys, has been downloaded 10 million times on Android devices across the world (according to the information available in the Google Play Store¹¹). People who actively participate in several panels are, however, suspected of satisficing and even forging data to increase their financial rewards (e.g., Toepoel et al., 2008; Yan & Tourangeau, 2008).

There is evidence that nonprobability survey respondents are a highly selective subgroup of the general population (e.g., Yeager et al., 2011). This is because people who recruit themselves by reacting to survey advertisements, which are a common way of nonprobability survey recruitment, are likely to have specific profiles in terms of demographics, personality, values, and habits. For instance, people who frequently use the Internet are more likely to be exposed to commercials and advertisements (e.g., OECD, 2014). In addition, previous research indicates that nonprobability survey participants show higher political knowledge and engagement (e.g., Chang & Krosnick, 2009; Duffy et al., 2005) and that they are unlikely to be older than 65 (e.g., Loosveldt & Sonck, 2009). We expect the self-recruitment procedure to negatively affect the representativeness of nonprobability

⁷ www.google.com/insights/consumersurveys/home

⁸ www.researchnow.com/audiences-data-collection/

⁹ www.omirussia.ru/en/online_panels/consumer_panels/#consumer

¹⁰ www.toluna-group.com/de/umfrage-plattform/toluna-samplexpress-

¹¹ play.google.com/store/apps/details?id=com.google.android.apps.paidtasks

survey data, even though there is some contradicting evidence (e.g., Gelman et al., 2017; Wang et al., 2015). Still, our first hypothesis is:

H1: Probability surveys are more representative than nonprobability surveys.

3.2 Response rates

Declining response rates are a widely-discussed issue in the survey methodological literature. In the German general social survey ALLBUS, for instance, response rates have been decreasing from 54% in 1994 to 38% in 2012 (Blohm & Koch, 2015). Similarly, Curtin et al. (2005) find a response rate decline of about one percentage point per year in the University of Michigan Survey of Consumer Attitudes. These findings are in accordance with earlier research on declining response rates (e.g., the international comparative study on household survey nonresponse by de Leeuw & de Heer, 2002). If less people participate in a survey, the risk of nonresponse bias increases (Bethlehem et al., 2011). This can be deduced from the well-known expression for nonresponse bias:

$$Bias(\bar{y}) = \frac{M}{N} (\bar{Y} - \bar{Y}_m) \quad (1)$$

“where $Bias(\bar{y})$ = the nonresponse bias of the unadjusted respondent mean; \bar{y}_r = the unadjusted mean of the respondents in a sample of the target population; \bar{Y}_r = the mean of the respondents in the target population; \bar{Y}_m = the mean of the nonrespondents in the target population; M = the number of nonrespondents in the target population; and N = the total number in the target population” (Groves, 2006: 648).

A difficulty with regard to response rates is that they are not always computed in the same way (for an overview, see e.g. AAPOR, 2016). In the scientific literature, it often remains unclear how exactly the response rates were calculated. This is especially true for the nonprobability surveys that usually provide participation rates where the number of the survey respondents is divided by the number of the invited members of the respondent pool instead of dividing the number of survey respondents by the number of persons in a sample, which is more common in the computation of probability

surveys. In the literature, these participation rates are often termed response rates although the AAPOR Task Force on Non-Probability Sampling, in an attempt to avoid confusion, recommends not calling these participation rates response rates (Baker et al., 2013).

Response rates are often reported as an indicator of nonresponse bias. It is, however, not necessarily true that high response rates lead to high degrees of survey representativeness. Indeed, the association between the response rate and nonresponse bias has been found to be small (Groves & Peytcheva, 2008). While it is true that the response rate places an upper bound on the potential nonresponse bias (Bethlehem, 2011), surveys with high response rates can still be heavily biased if the small number of nonrespondents systematically differs from the respondents. Conversely, surveys with relatively low response rates can accurately reflect population properties if the set of respondents only randomly varies compared to the set of nonrespondents.

Because the response rate places an upper bound on the potential nonresponse bias and because the response rate is often considered to be an indicator of the general quality of the survey data, we expect that surveys with high response rates more often achieve high degrees of representativeness. Therefore, our second hypothesis is:

H2: There is a positive association between the response rate and the measured degree of representativeness.

3.3 Mixed-mode surveys versus single-mode surveys

Mixed-mode surveys apply two or more modes of data collection (e.g. Dillman et al., 2008). Reasons for using multiple modes are diverse. One reason is that offline survey costs can be decreased by adding a web survey version (e.g., Couper et al., 2007; Jäckle et al., 2013). Another reason is that with different modes it is sometimes possible to reach a more diverse set of people and thereby survey representativeness increases (e.g., Lugtig et al., 2011; Klausch et al., 2015).

Mixed-mode survey designs vary in their purpose and type of implementation. Some surveys offer all respondents an option to choose their favorite mode from the onset (e.g., Bosnjak et al., 2017).

Other surveys assign modes to respondents based on estimated response propensities (e.g., Schouten & Cobben, 2007). Even more flexible mixed-mode designs are called ‘responsive design surveys’ (e.g., Groves & Heeringa, 2006; Peytchev et al., 2010). In responsive design studies, the survey mode assigned to a sample unit can change over the course of the survey fieldwork period depending on the development of subgroup response propensities. This approach has the advantage that modes can be targeted at sample subgroups that have a high risk of misrepresentation in the response set given one initial mode of data collection (e.g. singles in urban areas based on face-to-face survey data collection). These responsive techniques have become possible by the technological development in real-time fieldwork management (e.g., Macer, 2014; Blom, 2016).

Regardless of the specific design, mixed-mode surveys have in common that they strive to improve survey data quality as compared to traditional single-mode surveys. This goal might not always be achieved, especially when survey requests to an offline mode offer a concurrent web option (e.g., Medway & Fulton, 2012). Nevertheless, our third hypothesis is:

H3: Mixed-mode surveys are more representative than single-mode surveys.

3.4 Web surveys versus other single-mode surveys

Web surveys are a data collection mode that has become popular in survey research and practice in the early 2000s (e.g. Couper, 2000; Couper & Bosnjak, 2010). Since then, the number of web surveys has increased dramatically reaching turnovers of around 15 billion US\$ in 2014 in both Europe and North America¹². Generally, web surveys are a valuable addition to the possibilities of designing a survey. They are fast, cheap, and easy to conduct. In addition, they have the advantage that respondents can fill out questionnaires whenever and wherever they like by using mobile devices (e.g., Couper, 2013; Toepoel & Lugtig, 2015).

The rise of web surveys has, however, from the onset been accompanied by concerns about data quality, especially with regard to potential coverage error (e.g. Eckman, 2015; Blom et al., 2016). One

¹²www.esomar.org/news-and-multimedia/news.php?pages=1&idnews=150

reason for this is that web surveys are difficult to combine with probability sampling approaches, because often there is no sampling frame of Internet users available from which a sample could be drawn (e.g. Couper, 2000). Therefore, people who do not have an Internet connection and devices that enable logging into the Internet do not have the chance to be selected. In addition, people without access to the Internet have no chance to self-select into nonprobability web surveys. Research shows, however, that Internet users are systematically different from non-Internet users. For example, people with a low education level are usually underrepresented among Internet users (e.g., Blom et al., 2016).

Coverage error is more likely to be a problem in countries with relatively low Internet penetration rates than in countries with high Internet penetration rates (e.g., OECD, 2014). Bosnjak et al. (2013) find, however, that coverage error is the most influential source of attenuating representativeness in web surveys. Therefore, our fourth hypothesis is:

H4: Web surveys are less representative than other single-mode surveys.

3.5 Auxiliary variables

The term 'auxiliary data' usually encompasses all data that are available for both respondents and nonrespondents and can therefore be used to enhance post-survey adjustments (e.g., Kreuter, 2013). Examples of this type of data include sampling frame data, survey paradata, and data linked to survey data from external sources such as population registers. These data are often available on an aggregate level only, such as municipalities, city districts, or streets. Auxiliary data are commonly used to assess survey representativeness and adjust for misrepresentation (e.g., Brick & Kalton, 1996; Kreuter & Olson, 2011).

Some studies on survey representativeness only include basic sampling frame information or socio-demographic variables in the representativeness assessment because it is usually difficult to obtain more and other data for the assessment. Since these few basic variables are also commonly used to design quota samples or to construct nonresponse weights, many surveys cover these characteristics

sufficiently. However, the representativeness measure is generally more informative the broader the range of auxiliary variables taken into account actually is. The more auxiliary variables a study aims to address, the larger is the risk that at least one auxiliary variable is misrepresented in the data. For these reasons, our last hypothesis is:

H5: The more auxiliary variables are used for the representativeness assessment the lower is the overall representativeness.

4.Method

4.1 Literature search, study identification and data extraction

In order to answer our research question whether and how survey characteristics are related to survey representativeness, we have identified, reviewed, and coded the existing literature. In our dataset, we included all journal articles, book chapters, and scientific working papers that contain one of the two measures of representativeness that we focus on (general sample-based R-Indicators or descriptive comparisons between a survey and an external benchmark). Additional necessary preconditions for articles to be included in our analysis are that they need to be published in English, available in full text, and listed in an established database.

Our literature search and data coding were conducted in multiple steps by multiple coders. We used several data bases (Web of Science, EBSCO host, Jstor, and Google Scholar) and conducted full text searches wherever possible. Furthermore, we used the publications section of the website www.risq-project.eu as a database for articles on R-Indicators. Generally, we used a large number of search terms (“representativeness”, “representation”, “survey research”, “nonresponse”, and “R-Indicator”) in different combinations. We also conducted a snowballing search where we started with crucial articles and searched the literature section for relevant sources. For an overview of the article identification and eligibility assessment processes see Figure 1.

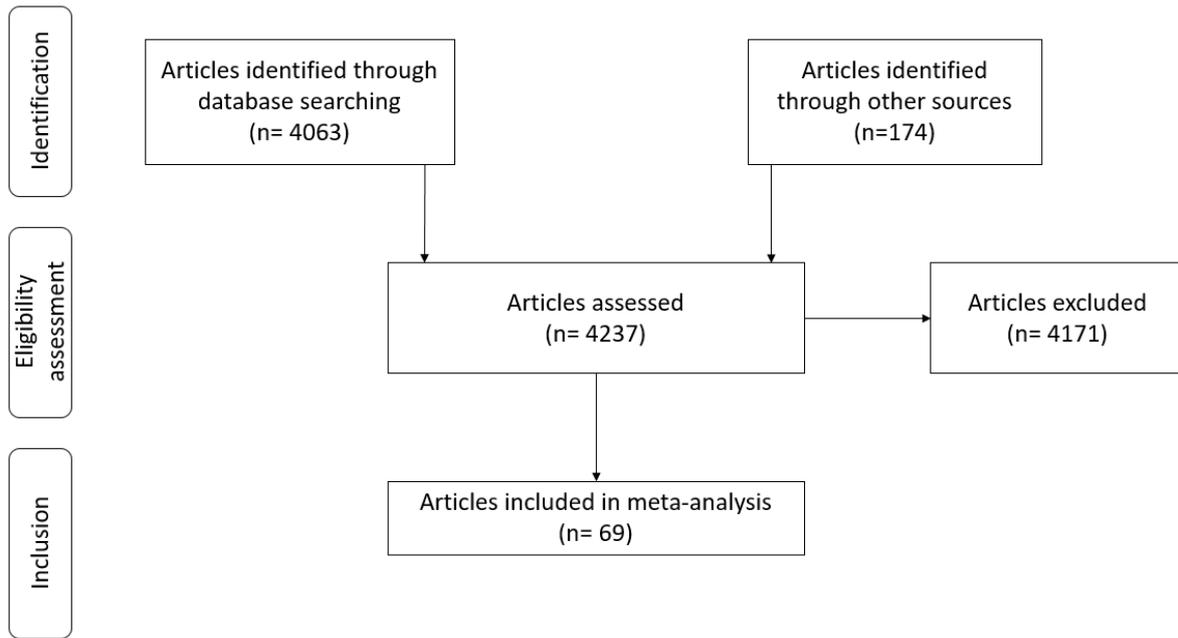


Figure 1: Article identification and eligibility assessment

As for the coding procedure, every data entry was checked by another coder and disagreements between coders were resolved by the authors of this paper. When the coded papers lacked necessary information (e.g., the survey mode), this information was searched on the Internet, requested from the authors of the papers in question via email, or, if possible, computed from the existing information.

4.2 Meta-analytic procedure

To compute the two overall mean representativeness scores across all eligible studies identified, we used Hedges/Olkin-type random-effects meta-analytic models (e.g., Hedges & Olkin, 1985; Raudenbush, 2009). In Hedges/Olkin-type meta-analyses, observed effect sizes (i.e., representativeness estimates in our case) are synthesized using a weighted mean procedure, with the inverse sampling variance of each effect size (i.e., representativeness coefficient) serving as weights. This procedure ensures that more precise representativeness estimates, that is, those being associated with smaller study-specific sampling variances, are assigned a larger weight when computing the overall mean representativeness estimate across all studies considered compared to the less precise effect sizes. Hedges/Olkin-type meta-analyses are based on one of two statistical models. A fixed-effect meta-analysis assumes all primary studies are estimating the same mean

(representativeness) value, with the only source of variability being study-level sampling error. A random-effects meta-analysis allows for unsystematic differences in the mean (representativeness) values from study to study, in addition to study-level sampling error. The selection of a model must be based on the question of which model fits the inferential goal, with random effects models being favoured for unconditional inferences, that is inferences going beyond the specific set of characteristics of the observed studies (Hedges & Vevea, 1998). Technically, in random effects meta-analyses, two sources of variability around the mean representativeness effect are estimated, namely unsystematic between-study variance ('T-square') in true effects, and within-study sampling error. As a next step, the homogeneity of the overall weighted representativeness means is estimated, answering the question if, and to what extent, the variability between observed representativeness measures can be explained by sampling error and T-square alone, and/or by systematic differences among effect sizes (so-called moderators). We report the Q statistic of the model that indicates heterogeneity if significant. In case of heterogeneity, which we assume in light of our hypotheses, we perform moderator analyses aiming to explain the variability among representativeness estimates using mixed-effects models that combine random-effects meta-analysis with data on the moderator. For an in-depth treatment of meta-analytic procedures, we recommend Borenstein et al. (2009), and Card (2011).

We use the `rma` function implemented in the R package *metafor*, version 2.0 (Viechtbauer, 2010). The `rma` function provides a general framework for fitting various meta-analytic models that are typically used in practice. Furthermore, we conduct moderator analyses using mixed-effects meta-regressions (van Houwelingen et al., 2002) implemented in *metafor*. The R-script and analysis output are available as Electronic Supplementary Material¹³.

4.3 Effect size measures

In meta-analytic terminology, dependent variables are called effect sizes and independent variables are called moderators. In the following, we describe the two effects sizes we use in this paper: R-

¹³<https://www.dropbox.com/sh/wmx1sap9e94fzto/AAABj7-H9ebkMS62WeTMPyqna?dl=0>

Indicators and the Median Absolute Bias (MAB) derived from the descriptive benchmark comparisons. These measures are common in the existing literature and examine survey representativeness from different perspectives.

4.3.1 R-Indicators

General sample-based R-Indicators are a measure of survey representativeness that is based on logistic regression models of the propensity to respond to a survey (Schouten et al., 2009). In practice, these response propensities are unknown and therefore have to be estimated, which is usually done using a logistic regression model. The independent variables in the regression models are auxiliary variables that need to be available for both respondents and nonrespondents to the survey. The individual response propensities from the regression model are aggregated using the formula

$$S(\rho) = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (\rho_i - \bar{\rho})^2} \quad (2)$$

Where n is the number of cases, i is the case indicator, ρ is the propensity to respond, and $\bar{\rho}$ is the mean response propensity.

The aggregated results from the propensity model are rescaled to range between zero (not representative) and one (very representative) using the formula

$$R(\rho) = 1 - 2S(\rho) \quad (3)$$

where $S(\rho)$ is the standard deviation of the response propensities. For our meta-regressions, we compute inverse variance weights for the R-Indicators based on their confidence intervals that we find in the publications. In the primary studies, they are usually computed from standard errors using bootstrapping procedures. Where confidence intervals for the R-Indicators are not reported we impute them using a predictive mean matching procedure (e.g., Schenker & Taylor, 1996).

4.3.2 MAB

We compute the Absolute Bias by category (AB_C) as the percentage point differences between proportions of a characteristic in a survey and an external gold-standard benchmark for each variable of a study with

$$AB_C = \left| \left(\frac{n_{R_C}}{n_R} \right) - \left(\frac{N_C}{N} \right) \right| \quad (4)$$

where n_R is the number of respondents, N is the benchmark estimate, and C is the characteristic.

The raw estimates of the survey and benchmark for our computations are usually reported in tables in the publications. To aggregate the results, we compute the Median Absolute Bias *MAB* across all of the characteristics assessed in a study of these individual percentage point differences across all categories and variables in the study. We compute inverse variance weights for the MABs based on bootstrapped standard errors (e.g., Efron & Tibshirani, 1986).¹⁴

5. Results

In this section, we summarize the meta-analytic findings. We start with some general results from random-effects meta-analyses, yielding a summary effect across all studies included for the two effect size measures considered, namely R-Indicators and MAB. Next, we give an overview of the moderator analyses conducted to test our five hypotheses, relating the following design features to representativeness: probability versus nonprobability surveys, response rates, mixed-mode surveys versus single-mode surveys, and the number of auxiliary variables. Finally, we present more detailed findings on each of our moderators in turn. In each of the analyses, significant outliers and missing values were excluded (see the Electronic Supplementary Material for detailed information).

¹⁴See the Electronic Supplementary Material for a sensitivity analysis using the Mean Absolute Bias instead of the Median Absolute Bias.

5.1 General findings

Generally, our random-effects models show for each of the two effect sizes (R-Indicators and MAB) that most surveys in our sample achieve high degrees of representativeness. In addition, we find that there is a substantial amount of heterogeneity in the data worth to be explored further.

Table 1: Random-effects meta-regression models of R-Indicators and MAB

	R-Indicators	MAB
Mean	0.84	4.39%
CI	0.82-0.85	3.73%-5.05%
Q-test	23286.35***	8710.95***
K	109	108

Note: Significance codes: 0 '***', 0.001 '**' 0.01 '*'

Table 1 displays the overall mean effect size, standard error, a heterogeneity estimate (Q-test results), and the number of studies (k) included in our analysis of the random-effects models on R-Indicators and the MAB. The mean effect sizes are 0.84 for the R-Indicators and 4.39% for the MAB. The confidence intervals (CI) around these mean effect sizes are from 0.82 to 0.85 for the R-Indicators and from 3.73% to 5.05% for the MAB. The Q-tests are highly significant for both effect sizes, indicating substantial heterogeneity in the data worth exploring with the aid of moderator analyses.

5.2 Moderator analyses

Overall, our moderator analyses show that there are significant associations between the degree of representativeness as measured using R-Indicators or the MAB and the survey design characteristics specified in our hypotheses. In Table 2, we display an overview of the findings. The table is structured as follows: The first column contains the names of the moderators and information on their variable type (dichotomous versus continuous). The estimates of the continuous variables were rescaled for easier interpretation of the results. The second column lists the categories of the dichotomous variables. The rest of the table is divided into an overview of results from the moderator analyses on the R-Indicators and those of the MAB. For each of the effect sizes, we present the number of studies included in the moderator analyses (k). If the moderator is a dichotomous variable, we display the

number of studies included for each of the moderator's categories separately. Furthermore, we present the regression coefficients, their significance level, and their standard errors of the mixed-effects meta-regressions of the effect sizes on the individual moderators as well as R²-statistics as a measure of model fit.

Table 2: Mixed-effects meta-regression models of R-Indicators on each moderator (standard errors in parentheses)

Moderator (Variable type) Categories		R-Indicator			MAB		
		k	Coefficient	R ²	k	Coefficient	R ²
Sample (dichotomous)	Prob.	110	-	-	61	-2.18***	10.93%
	Nonprob.	0	-		49	(0.59)	
Response rate (continuous; rescaled x100)		104	0.14** (0.04)	8.18%	90	-2.05* (1.14)	2.22%
Mixed mode (dichotomous)	Mixed	45	0.05***	13.22%	8	-	-
	Single	51	(0.01)		101	-	
Web survey (dichotomous)	Web	1	-	-	56	1.57*	4.37%
	Other	50	-		45	(0.65)	
No. auxiliary variables (continuous; rescaled x100)		104	0.05 (0.00)	0.00%	104	-10.28* (4.48)	3.81%

Note: Significance codes: 0 '***', 0.001 '**' 0.01 '*'

Generally, we find in the mixed-effects meta-regression models that the degree of representativeness as measured using R-Indicators is associated significantly positively with the response rate and with mixed-mode surveys (as opposed to single-mode surveys). In addition, the MAB is significantly negatively associated with probability surveys (as opposed to nonprobability surveys), the response rate, other single-mode surveys (as opposed to web surveys), and the number of auxiliary variables on which representativeness is assessed. Next, we discuss the results of the analyses on each moderator in more detail.

5.2.1 Probability versus nonprobability surveys

The first moderator we examine is probability surveys versus nonprobability surveys. Overall, we find evidence in support of our expectation that probability surveys are more representative than

nonprobability surveys (H1), although we can only identify this association using the MAB as a representativeness measure.

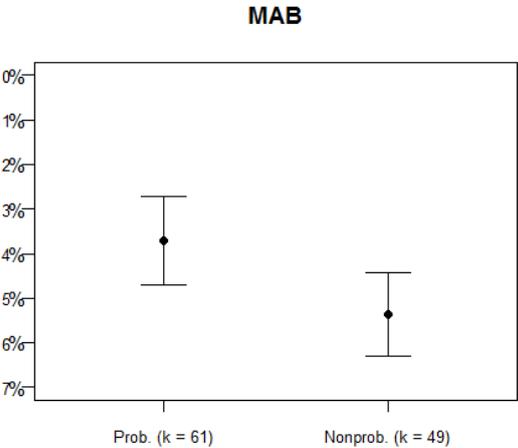


Figure 2: Subgroup comparison results by probability versus nonprobability surveys as a moderator

The computation of sample-based R-Indicators requires that a survey is based on a probability sample. In nonprobability samples, we are unable to identify the nonrespondents, and we also do not have auxiliary data on the nonrespondents. A workaround might be the usage of population-based R-Indicators, which apply population aggregated auxiliary information, for example from benchmark data, instead of individual-level auxiliary data (e.g., Shlomo et al., 2009). These population-based R-Indicators are not yet widely used, especially not with regard to nonprobability surveys. We therefore cannot consider them in our analysis. Therefore, we do not have any nonprobability surveys in the R-Indicator analysis that we could compare the probability surveys to.

Regarding the MAB, we find that there are 61 probability surveys and 49 nonprobability surveys in our dataset. The right-hand side of Figure 2 shows an average MAB of 3.72% with a confidence interval from 2.72% to 4.71% in probability surveys and an MAB of 5.36% with a 95%-confidence interval from 4.42% to 6.30% in nonprobability surveys. Table 2 shows that the regression coefficient from the mixed-effects meta-regression model is -2.18 and highly significant, the R²-statistic for this model is 10.93%. These findings suggest that probability surveys are more representative than nonprobability surveys, which is in accordance with our expectations.

5.2.2 Response rates

The second moderator we examine is the response rate, which is a continuous variable. Overall, we find that the response rate is highly significantly associated with the degree of representativeness as measured using R-Indicators and the MAB. This is in accordance with our second expectation (H2).

As displayed in Table 2, the number of R-Indicators in our mixed-effects meta-regression model of the response rate on the R-Indicators is 104. The model's rescaled coefficient is 0.14, significant, and has a standard error of 0.04%. The R^2 -statistic of this model is 8.18%. The number of MAB studies that we use in our moderator analysis on the response rate is 90. The mixed-effects meta-regression produces a rescaled coefficient of -2.05 that is significant and has a standard error of 1.14 as well as an R^2 -statistic of 2.22%.

Overall, we find across R-Indicators and the MAB that the results of our meta-regressions support our expectation that response rates are positively associated with the reported degree of representativeness.

5.2.3 Mixed-mode versus single-mode surveys

The third moderator we investigate is a dichotomous variable on mixed-mode surveys versus single-mode surveys. Our results suggest that mixed-mode surveys are more representative than single-mode surveys (H3) although there are too few mixed-mode surveys (8 surveys) among the MAB studies to allow for valid conclusions.

There are 45 mixed-mode surveys and 51 single-mode surveys among the R-Indicators in our dataset. Figure 3 shows that the average R-Indicators are 0.87 with a confidence interval from 0.85 to 0.89 in the mixed-mode surveys and 0.83 with a confidence interval from 0.82 to 0.84 in the single-mode surveys. The mixed-effects meta-regression reported in Table 2 has a coefficient of 0.04 with a standard error of 0.01 and an R^2 -statistic of 13.22%.

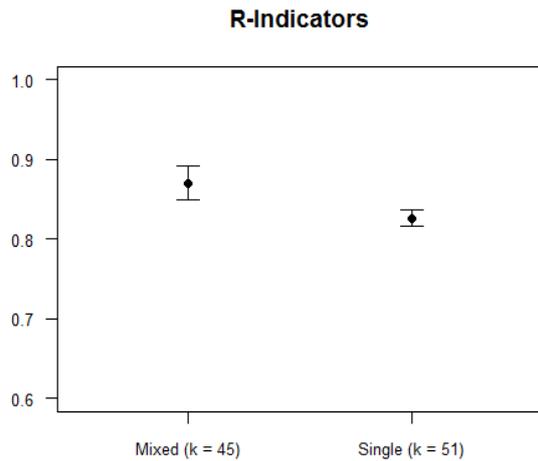


Figure 3: Subgroup comparison results by mixed-mode versus single-mode surveys as a moderator

The number of MAB studies is 8 on mixed-mode surveys and 101 on single mode surveys. Because of the insufficient number of cases in the mixed-mode survey subgroup, estimation of a mixed-effects meta-regression model on this moderator is not feasible.

Overall, we find support for our expectation that mixed-mode surveys are more representative than single-mode surveys on the R-Indicators while for the MAB, the available evidence at present does not allow us to conduct moderator analyses.

5.2.4 Web surveys versus other single-mode surveys

Next, we investigate the association between survey representativeness and web surveys versus other single-mode surveys as a moderator. In accordance with our expectation (H4), our findings indicate, that web surveys are less representative than single-mode surveys as measured using the MAB. There is only one R-Indicator study that examines web surveys. Therefore we cannot conduct a moderator analysis on the R-Indicator data.

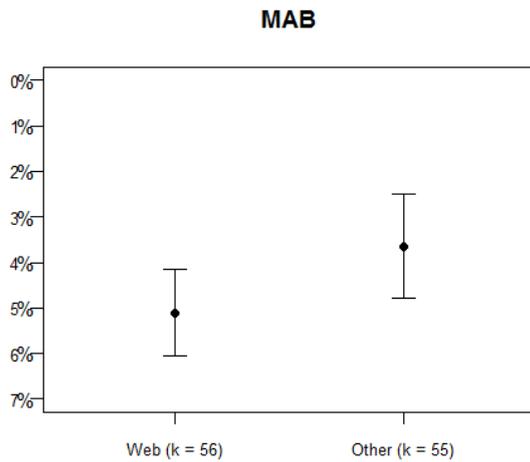


Figure 4: Subgroup comparison results by web surveys versus other single-mode surveys as a moderator

The number of R-Indicator studies is 1 for the web surveys and 50 for other single-mode surveys. The remaining R-Indicator studies identified do not cover single-mode surveys. Because there is only one web survey in our R-Indicator data, we do not estimate a mixed-effects meta-regression model on this moderator.

There are 56 web surveys and 45 other single-mode surveys in our MAB data. The average MAB is 5.10% with a confidence interval from 4.15% to 6.06% in the web surveys and 3.65% with a confidence interval from 2.51% to 4.80% in the other single mode surveys. The mixed-effects meta-regression on this moderator estimates a regression coefficient of 1.57 that is significant and has a standard error of 0.65. The R^2 of this model is 4.37%.

In sum, the MAB data support our expectation that web surveys are less representative than other single-mode surveys while there is only one web survey among the R-Indicator studies in our data so that we cannot draw valid conclusions on this moderator for R-Indicators.

5.2.5 Number of auxiliary variables

The last moderator we examine is the number of auxiliary variables that enters the representativeness assessment. Overall, we find no evidence in favor of our expectation (H5). This suggests that there is no negative association between the number of auxiliary variables included in a representativeness assessment and the reported degree of representativeness.

There are 104 R-Indicator studies that we use in the moderator analysis concerning the number of auxiliary variables. The rescaled regression coefficient in the mixed-effects meta-regression model is 0.05, not significant, and has a standard error of 0.00. The R^2 of the model is 0.00%. Regarding the MAB, there are 104 studies included in the moderator analysis of the number of auxiliary variables. The rescaled mixed-effects meta-regression coefficient is -10.28, significant, and has a standard error of 4.48. The R^2 -statistic of this model is 3.81%.

Overall, the findings on this moderator are not entirely robust across the two effect sizes. However, our findings show that neither on the R-Indicators nor on the MAB there is a negative association between the number of auxiliary variables included in a representativeness assessment and the reported degree of representativeness. In fact, the MAB findings indicate that more auxiliary variables are positively associated with reported representativeness.

6. Summary and conclusion

In this section, we revisit our expectations from above and interpret the empirical findings. Then, we discuss practical implications and limitations of our paper as well as avenues for further research.

In accordance with our first expectation (H1), we find that probability surveys are more representative than nonprobability surveys with regard to the MAB. In line with our second expectation (H2), we find that the response rate is positively associated with representativeness on R-Indicators as well as the MAB. In compliance with our third expectation (H3), we find that mixed-mode surveys are more representative than single-mode surveys on the R-Indicators. Furthermore, in accordance with our fourth expectation (H4), we find that web surveys are less representative than other single-mode surveys with regard to the MAB. Contrary to our fifth expectation (H5) we find that there is no negative association between the number of auxiliary variables and the reported degree of representativeness.

We expected results to be consistent across the two measures of representativeness. Due to insufficient numbers of cases per category, however, some expectations could only be tested for one

of the two measures. There are two moderator analyses (on the response rate and the number of auxiliary variables) that we could conduct on both representativeness measures. Both the R-Indicators and the MAB indicate that the higher the response rate is the higher is the reported degree of representativeness. In addition, we find on both the R-Indicators and the MAB that there is no negative association between the number of auxiliary variables and reported representativeness. However, only the MAB indicates that in fact the association between the number of auxiliary variables and the degree of representativeness might be positive.

Based on these findings and disregarding potential interdependencies between moderator variables in our data that might confound our results, we would recommend survey practitioners aiming to design representative surveys to use probability sampling, to put every effort into increasing response rates, and to consider mixed-mode survey designs encompassing the web mode. We would also recommend documenting the achieved degree of representativeness using the available auxiliary data to allow assessing the general quality of the data and to allow further research into the association between survey design characteristics and survey representativeness.

This study does, however, face some restrictions that limit the generalizability of the results. Firstly, the analyses do not allow for causal inferences. We cannot claim that probability sampling, high response rates, mixed-mode designs, and the other-than-web survey modes cause a survey to be representative. Our analyses are limited to identifying existing associations. These associations might be confounded with other survey characteristics that we cannot control for in our analyses, such as the data quality of the auxiliary data used in the primary research. Secondly, since our analyses are based on the existing literature, publication biases might influence our results (see the Electronic Supplementary Material for sensitivity analyses). This is especially the case when researchers do not publish results that show that a survey has a low degree of representativeness. Lastly, and potentially partly due to publication bias, the number of cases on some survey characteristics is too small to conduct all planned analyses. In the existing literature, there are few studies that assess the representativeness of mixed-mode surveys using benchmark comparisons that allow the

computation of the MAB or that investigate the representativeness of web surveys using R-Indicators. In addition, there are no studies that assess the representativeness of a survey using R-Indicators as well as benchmark comparisons. Furthermore, there are a number of potentially influential moderators that we could not take into account either because too much information was missing from the publications identified, or the information was too ambiguous, for instance regarding whether and how post-survey adjustments were applied.

We would therefore like our recommendations to be taken with caution and this meta-analysis to be considered as encouragement for more systematic primary research in these areas, filling the gaps identified. If there were more primary research available, future meta-analytic replications could be conducted using more advanced modeling, such as multivariate regression analyses, simultaneously controlling for potentially influential moderators. In a future replication and extension of this meta-analysis, besides considering additional primary studies, one might want to add additional moderators of practical importance, such as whether surveys apply a responsive mode design, which formula is used to compute the response rate, or whether adjustment weights were applied.

Literature

American Association for Public Opinion Research (AAPOR) (2016). *Standard Definitions: Final Dispositions of Case Codes and Outcome Rates for Surveys*. 9th edition. Retrieved from http://www.aapor.org/AAPOR_Main/media/publications/Standard-Definitions20169theditionfinal.pdf.

Baker, R., Brick, M. J., Bates, N., Battaglia, M., Couper, M. P., Dever, J. A., Gile, K. J., Tourangeau, R. (2013). *Report of the AAPOR Task Force on Non-Probability Sampling*. Retrieved from https://www.aapor.org/AAPOR_Main/media/MainSiteFiles/NPS_TF_Report_Final_7_revised_FNL_6_22_13.pdf.

Bethlehem, J., & Biffignandi, S. (2012). *Handbook of Web Surveys*. New York, NY: Wiley.

Bethlehem, J., Cobben, F., & Schouten, B. (2011). *Handbook of Nonresponse in Household Surveys*. Hoboken, NJ: John Wiley & Sons.

Blohm, M., & Koch, A. (2015). Führt eine höhere Ausschöpfung zu anderen Umfrageergebnissen? Eine experimentelle Studie zum ALLBUS 2008. In J. Schupp & C. Wolf (Eds.) *Nonresponse Bias. Qualitätssicherung sozialwissenschaftlicher Umfragen* (pp. 85–129). Wiesbaden: Springer Fachmedien.

Blom, A. G. (2016). Survey Fieldwork. In C. Wolf, D. Joye, T. W. Smith & Y. Fu (Eds.) *The SAGE Handbook of Survey Methodology* (pp. 382–397). London, UK: Sage Publications.

Blom, A. G., Herzing, J.M.E., Cornesse, C., Sakshaug, J. W., Krieger, U., & Bossert, D. (2016). Does the Recruitment of Offline Households Increase the Sample Representativeness of Probability-Based Online Panels? Evidence From the German Internet Panel. *Social Science Computer Review*, 1(1), 1–23.

Borenstein, M., Hedges, L. V., Higgins, J.P.T., & Rothstein, H. R. (2009). *Introduction to Meta-Analysis*. New York, NY: Wiley.

Bosnjak, M. (2017). Mixed-Mode Surveys and Data Quality. Meta-Analytic Evidence and Avenues for Future Research. In S. Eifler & F. Faulbaum (Eds.), *Methodische Probleme von Mixed-Mode-Ansätzen in der Umfrageforschung* (pp. 11–25). Wiesbaden: Springer Fachmedien.

Bosnjak, M., Dannwolf, T., Enderle, T., Schaurer, I., Struminskaya, B., Tanner, A., & Weyandt, K.W. (2017). Establishing an open probability-based mixed-mode panel of the general population in Germany: The GESIS Panel. *Social Science Computer Review*.

Bosnjak, M., Haas, I., Galesic, M., Kaczmirek, L., Bandilla, W., & Couper, M. P. (2013). Sample composition biases in different development stages of a probabilistic online panel. *Fields Methods*, 25(4), 339–360.

Brick, M. J., & Kalton, G. (1996). Handling missing data in survey research. *Statistical Methods in Medical Research*, 5(1), 215–238.

Brick, M. J., & Williams, D. (2013). Explaining Rising Nonresponse Rates in Cross-Sectional Surveys. *The Annals of the American Academy of Political and Social Science*, 645(1), 36–59.

Card, N. A. (2012). *Applied Meta-Analysis for Social Science Research*. New York, London: The Guilford Press.

Chang, L., & Krosnick, J. A. (2009). National Surveys Via Rdd Telephone Interviewing Versus the Internet: Comparing Sample Representativeness and Response Quality. *Public Opinion Quarterly*, 73(4), 641–678.

Couper, M.P. (2000). Web Surveys: A Review of Issues and Approaches. *Public Opinion Quarterly*, 64(1), 464–494.

Couper, M.P. (2013). Is the Sky Falling? New Technology, Changing Media and the Future of Surveys. *Survey Research Methods*, 7(3), 145–156.

Couper, M. P., & Bosnjak, M. (2010). Internet Surveys. In P. V. Marsden & J. D. Wright (Eds.) *Handbook of Survey Research* (pp. 527–550). San Diego, CA: Elsevier.

- Couper, M.P., Kapteyn, A., Schonlau, M., & Winter, J. (2007). Noncoverage and nonresponse in an Internet survey. *Social Science Research, 36*(1), 131–148.
- Curtin, R., Presser, S., & Singer, E. (2005). Changes in Telephone Survey Nonresponse over the Past Quarter Century. *Public Opinion Quarterly, 69*(1), 87–98.
- de Leeuw, E. D., & de Heer, W. (2002). Trends in household survey nonresponse: A longitudinal and international comparison. In R. M. Groves, A. D. Dillman, J. Eltinge, R. Little (Eds.), *Survey nonresponse* (pp. 41–54.). New York, NY: John Wiley & Sons.
- Dillman, A. D. (2008). The logic and psychology of constructing questionnaires. In E. D. de Leeuw, J. J. Hox, A. D. Dillman (Eds.), *International handbook of survey methodology* (pp. 161–175). New York, NY: Lawrence Erlbaum Associates.
- Duffy, B., Smith, K., Terhanian, G., & Bremer, J. (2005). Comparing Data from Online and Face-to-Face Surveys. *International Journal of Market Research, 47*(6), 615–639.
- Eckman, S. (2015). Does the Inclusion of Non-Internet Households in a Web Panel Reduce Coverage Bias? *Social Science Computer Review, 34*(1), 41–58.
- Efron, B., & Tibshirani, R. J. (1986). Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statistical Science, 1*(1): 54–77.
- Emerson, M. O., Sikkink, D., & James, A. D. (2010). The Panel Study on American Religion and Ethnicity: Background, Methods, and Selected Results. *Journal for the Scientific Study of Religion, 49*(1), 162–171.
- Gelman, A., Goel, S., Rothschild, D., & Wang, W. (2017). High-frequency polling with non-representative data. In D. Schill, R. Kirk, A. E. Jasperson (Eds.) *Political Communication in Real Time. Theoretical and Applied Research Approaches* (pp. 89-105). New York, NY: Routledge.
- Goldenbeld, C., & Craen, S. (2013). The comparison of road safety survey answers between web-panel and face-to-face: Dutch results of SARTRE-4 survey. *Journal of Safety Research, 46*(1), 13–20.

- Göriz, A. S. (2004). The impact of material incentives on response quantity, response quality, sample composition, survey outcome, and cost in online access panels. *International Journal of Market Research*, 46(3), 327–345.
- Groves, R. M. (2006). Nonresponse rates and nonresponse bias in household surveys. *Public Opinion Quarterly*, 70(5), 646–675.
- Groves, R. M., Fowler, F. J., Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2004). *Survey Methodology*. Hoboken, NJ: John Wiley & Sons.
- Groves, R. M., Brick, J. M., Couper, M.P, Kalsbeek, W., Harris-Kojetin, B., Kreuter, F., Pennell, B.-E., Raghunathan, T., Schouten, B., Smith, T., Tourangeau, R., Bowers, A., Jans, M., Kennedy, C., Levenstein, R., Olson, K., Peytcheva, E., Ziniel, S., Wagner, J. (2008). Issues Facing the Field: Alternative Practical Measures of Representativeness of Survey Respondent Pools. *Survey Practice*, 1(3), 1–6.
- Groves, R. M., & Heeringa, S. G. (2006). Responsive design for household surveys: tools for actively controlling survey errors and costs. *Journal of the Royal Statistical Society: Series A*, 169(3), 439–457.
- Groves, R. M., & Lyberg, L. (2010). Total survey error: Past, present, and future. *Public Opinion Quarterly*, 74(5), 849–879.
- Groves, R. M., & Peytcheva, E. (2008). The impact of nonresponse rates on nonresponse bias. *Public Opinion Quarterly*, 72(2), 167–189.
- Hedges, L. V., & Olkin, I. (1985). *Statistical Methods for Meta-Analysis*. London, UK: Academic Press.
- Hedges, L. V., & Vevea, J. L. (1998). Fixed-and random-effects models in meta-analysis. *Psychological Methods*, 3(4), 486-504.
- Jardine, C. G., Boerner, F. U., Boyd, A. D., & Driedger, S. M. (2015). The More the Better? A Comparison of the Information Sources Used by the Public during Two Infectious Disease Outbreaks. *PLoS ONE*, 10(10): e0140028.

- Jäckle, A., Lynn, P., & Burton, J. (2013). Going Online with a Face-to-Face Household panel: Initial Results from an Experiment on the Understanding Society Innovation Panel. *Economic and Social Research Council*, Colchester, UK.
- Jäckle, A., Roberts, & C., Lynn, P. (2010). Assessing the Effect of Data Collection Mode on Measurement. *International Statistical Review*, 78(1), 3–20.
- Keeter, S., Miller, C., Kohut, A., Groves, R. M. & Presser, S. (2000). Consequences of Reducing Nonresponse in a National Telephone Survey. *Public Opinion Quarterly*, 64(1), 125–148.
- Kish, L. (1965). *Survey Sampling*. New York, NY: John Wiley & Sons.
- Klausch, T., Hox, J. J., & Schouten, B. (2015). Selection error in single- and mixed mode surveys of the Dutch general population. *Journal of the Royal Statistical Society: Series A*, 178(4), 945–961.
- Kreuter, F. (2013). *Improving Surveys with Paradata*. New York, NY: Wiley.
- Kreuter, F., & Olson, K. (2011). Multiple auxiliary variables in nonresponse adjustment. *Sociological Methods & Research*, 40(2), 311–332.
- Kruskal, W., & Mosteller, F. (1979a). Representative Sampling, I: Non-Scientific Literature. *International Statistical Review*, 47(1), 13–24.
- Kruskal, W., & Mosteller, F. (1979b). Representative Sampling, II: Scientific Literature, Excluding Statistics. *International Statistical Review*, 47(2), 111–127.
- Kruskal, W., & Mosteller, F. (1979c). Representative Sampling, III: The Current Statistical Literature. *International Statistical Review*, 47(3), 245–265.
- Lohr, S. (2010). *Sampling: Design and Analysis*. Brooks/Cole: Cengage Learning.
- Loosveldt, G., & Sonck, N. (2008). An evaluation of the weighting procedures for an online access panel survey. *Survey Research Methods*, 2(2), 93–105.

- Lugtig, P., Lensvelt-Mulders, G. J. L. M., Frerichs, R., & Greven, A. (2011). Estimating nonresponse bias and mode effects in a mixed-mode survey. *International Journal of Market Research*, 53(5), 2–16.
- Luiten, A., & Schouten, B. (2012). Tailored fieldwork design to increase representative household survey response: an experiment in the Survey of Consumer Satisfaction. *Journal of the Royal Statistical Society (Series A)*, 176(1), 169–189.
- Lundquist, P., & Särndal, C.E. (2013). Aspects of Responsive Design with Applications to the Swedish Living Conditions Survey. *Journal of Official Statistics*, 29(4), 557–582.
- Macer, T. (2014). Online panel software. 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. 413–438). London, UK: John Wiley & Sons.
- Malhotra, N., & Krosnick, J. A. (2007). The Effect of Survey Mode and Sampling on Inferences about Political Attitudes and Behavior: Comparing the 2000 and 2004 ANES to Internet Surveys with Nonprobability Samples. *Political Analysis*, 15(3), 286–323.
- Medway, R. L., & Fulton, J. (2012). When More Gets You Less: A Meta-Analysis of the Effect of Concurrent Web Options on Mail Survey Response Rates. *Public Opinion Quarterly*, 76(4), 733–746.
- OECD (2014). *Measuring the Digital Economy: A New Perspective*. OECD Publishing.
- Peytchev, A., Conrad, F.G., Couper, M.P., & Tourangeau, R. (2010). Increasing Respondents' Use of Definitions in Web Surveys. *Journal of Official Statistics*, 26(4), 633–650.
- Potthof, P., Heinemann, L. A. J., & Güther, B. (2004). A Household Panel as a Tool for Cost-Effective Health Related Population Surveys: Validity of the "Healthcare Access Panel". *German Medical Science*, 2(1), 1–8.
- Raudenbush, S. W. (2009). Analyzing effect sizes: Random effects models. In H. Cooper, L. V. Hedges, C. Valentine (Eds.) *The handbook of research synthesis and meta-analysis* (pp. 295–315). New York, NY: Russell Sage Foundation.

- Rendtel, U., & Pötter, U. (1992). Über Sinn und Unsinn von Repräsentationsstudien. *DIW Discussion Papers*, 61.
- Revilla, M., Cornilleau, A., Cousteaux, A. S., Legleye, S., & Pedraza, P. (2016). What Is the Gain in a Probability-Based Online Panel of Providing Internet Access to Sampling Units Who Previously Had No Access? *Social Science Computer Review*, 34(4), 479–496.
- Schenker, N., & Taylor, J. M. G. (1996). Partially parametric techniques for multiple imputation. *Computational Statistics & Data Analysis*, 22, 425-446.
- Schnell, R. (1993). Die Homogenität sozialer Kategorien als Voraussetzung für „Repräsentativität“ und Gewichtungungsverfahren. *Zeitschrift für Soziologie*, 22(1), 16–32.
- Schouten, B., & Cobben, F. (2007). R-indexes for the comparison of different fieldwork strategies and data collection modes. *Statistics Netherlands: Discussion paper 07002*.
- Schouten, B., Cobben, F., Bethlehem, J. (2009). Indicators for the representativeness of survey response. *Survey Methodology*, 35(1), 101–113.
- Shlomo, N., Skinner, C., Schouten, B., de Heij, V., Bethlehem, J., & Ouwehand, P. (2009). *Indicators for Representative Response Based on Population Totals*. RISQ Work package 3, Deliverable 2.2. Retrieved from <http://www.risq-project.eu/papers/RISQ-Deliverable-2-2-V1.pdf>
- Sterrett, D., Malato, D., Benz, J., Tompson, T., & English, N. (2017). Assessing Changes in Coverage Bias of Web Surveys in the United States. *Public Opinion Quarterly*, 81(Special Issue), 338-356.
- Toepoel, V., Das, M., & van Soest, A. (2008). Effects of design in web surveys comparing trained and fresh respondents. *Public Opinion Quarterly*, 72(5), 985–1007.
- Toepoel, V., & Lugtig, P. (2015). Online Surveys are Mixed-Device Surveys. Issues Associated with the Use of Different (Mobile) Devices in Web Surveys. *Methods, data, analyses*, 9(2), 155–162.
- Tourangeau, R., Conrad, F.G., & Couper, M.P. (2013). *The Science of Web Surveys*. Oxford: Oxford University Press.

- van Houwelingen, H. C., Arends, L. R., & Stijnen, T. (2002). Advanced methods in meta-analysis: Multivariate approach and meta-regression. *Statistics in Medicine*, 21(1), 589–624.
- Viechtbauer, W. (2010). Conducting Meta-Analyses in R with the metafor Package. *Journal of Statistical Software*, 36(2), 1–48.
- Vonk, T., van Ossenbruggen, R., & Willems, P. (2006). The effects of panel recruitment and management on research results, a study among 19 online panels. *ESOMAR Publication Services*, 317, 79–99.
- Wagner, J. (2010). The Fraction of Missing Information as a Tool for Monitoring the Quality of Survey Data. *Public Opinion Quarterly*, 74(2), 223–243.
- Wagner, J. (2012). Research Synthesis. A Comparison of Alternative Indicators for the Risk of Nonresponse Bias. *Public Opinion Quarterly*, 76(3), 555–575.
- Wang, W., Rothschild, D., Goel, S., & Gelman, A. (2015). Forecasting elections with non-representative polls. *International Journal of Forecasting*, 31(1), 980–991.
- Wozniak, K. H. (2016). Perceptions of Prison and Punitive Attitudes: A Test of the Penal Escalation Hypothesis. *Criminal Justice Review*, 1(1), 1–20.
- 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.
- Yeager, D. S., Krosnick, J. A., Chang, L., Lavitz, H. S., Levendusky, M. S., Simpser, A., & Wang, R. (2011). Comparing the accuracy of RDD telephone surveys and internet surveys conducted with probability and non-probability samples. *Public Opinion Quarterly*, 75(4), 709–747.

Online Appendix A: R-Indicator articles in the meta-analysis

Aramendi, J., Goni, E., & Iztueta, A. (2012, May, 30). *Non response measurement: R-Indicators and further treatments in Eustat*. Paper presented at the European Conference on Quality in Official Statistics (Q2012), Athens, Greece. Retrieved from <http://www.q2012.gr/articlefiles/sessions/30.2-Q2012.pdf>.

Ariel, A., & Schouten, B. (2008). Representativity of the Time Use Survey. Statistics Netherlands: *Discussion paper (08013)*. Statistics Netherlands: Voorburg/Heerlen.

Beullens, K., & Loosveldt, G. (2010, May, 4-6). *R-indicators and Fieldwork Monitoring*. Paper presented at the European Conference on Quality in Official Statistics (Q2010), Helsinki, Finland.

Bethlehem, J. G. (2010). New developments in survey data collection methodology for official statistics. *Discussion paper (10010)*. Statistics Netherlands: Voorburg/Heerlen.

Bethlehem, J. G., Cobben, F., & Schouten, B. (2011). A New Quality Indicator for Survey Response. *AENORM, 19(72)*, 24–28.

Cobben, F., & Schouten, B. (2008). An empirical validation of R-indicators. *Discussion paper (08006)*. Statistics Netherlands: Voorburg/Heerlen.

Keathley, D., & Hefter, S. (2013). *American Community Survey: Sample Representivity for the Nation and Puerto Rico. Memorandum for ACS Research Evaluation Advisory Group*. Retrieved from <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.406.4745&rep=rep1&type=pdf>.

Luiten, A., & Schouten, B. (2012). Tailored fieldwork design to increase representative household survey response: an experiment in the Survey of Consumer Satisfaction. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 176(1), 169-189.

Lundquist, P., & Särndal, C.E. (2013). Aspects of Responsive Design with Applications to the Swedish Living Conditions Survey. *Journal of Official Statistics*, 29(4), 557–582.

- Nooij, G. de (2008). Representativity of Short Term Statistics. *Afstudeerscriptie*. The Netherlands: Vrije Universiteit. Retrieved from http://www.few.vu.nl/nl/Images/stageverslag-nooij_tcm243-90671.pdf.
- Ouwehand, P., & Schouten, B. (2014). Measuring Representativeness of Short-Term Business Statistics. *Journal of Official Statistics*, 30(4), 623–649.
- Roberts, C., Vandenplas, C., & Stähli, M. E. (2014). Evaluating the Impact of Response Enhancement Methods on the Risk of Nonresponse Bias and Survey costs. *Survey Research Methods*, 8(2), 67–80.
- Schouten, B., Cobben, F., & Bethlehem, J. (2009). Indicators for the representativeness of survey response. *Survey Methodology*, 35(1), 101–113.
- Schouten, B., & Cobben, F. (2012). Does balancing of survey response reduce nonresponse bias? *Discussion paper (201226)*. Statistics Netherlands: Voorburg/Heerlen. Retrieved from <https://www.cbs.nl/-/media/imported/documents/2014/04/2012-26-x10-pub.pdf>.
- Schouten, B., Shlomo, N., & Skinner, C. (2010, May, 4-6). *Indicators for Representative Response*. Paper presented at the European Conference on Quality in Official Statistics (Q2010), Helsinki, Finland.
- Schouten, B., Shlomo, N., & Skinner, C. (2011). Indicators for Monitoring and Improving Representativeness of Response. *Journal of Official Statistics*, 27(2), 1–24.
- Schouten, B., Bethlehem, J., Beullens, K., Kleven, Ø., Loosveldt, G., Luiten, A., Rutar, K., Shlomo, N., & Skinner, C. (2012). Evaluating, Comparing, Monitoring, and Improving Representativeness of Survey Response Through R-Indicators and Partial R-Indicators. *International Statistical Review*, 80(3), 382–399.
- Shlomo, N., Skinner, C., & Schouten, B. (2012). Estimation of an indicator of the representativeness of survey response. *Journal of Statistical Planning and Inference*, 142(1), 201–211.

Shlomo, N., Schouten, B., & Heij, V. de (2013, March, 5-7). *Designing Adaptive Survey Designs with R-Indicators*. Paper presented at the New Techniques and Technologies for Statistics (NTTS) Conference, Brussels, Belgium.

Shlomo, N., Skinner, C., Schouten, B., Bethlehem, J. G. & Zhang, L. C. (2009). Statistical Properties of R-Indicators, Version 2. *Representativity Indicators for Survey Quality*. Retrieved from <http://risq-project.eu/papers/RISQ-Deliverable-2-1-V2.pdf>.

Skinner, C., Shlomo, N., Schouten, B., Zhang, L. C., & Bethlehem, J. G. (2009, February, 18-20). *Measuring Survey Quality Through Representativeness Indicators Using Sample and Population Based Information*. Paper presented at the New Techniques and Technologies for Statistics (NTTS) Conference, Brussels, Belgium. Retrieved from <http://www.risq-project.eu/papers/skinner-shlomo-schouten-zhang-bethlehem-2009-a.pdf>.

Whitehead, D., Broderick, O., & Gonzalez, Y. (2014, August, 2-7). *The Use of Indicators to Assess the Quality of Business Survey Returns During Data Collection*. Paper presented at the Joint Statistical Meeting, Boston, Massachusetts. Retrieved from http://ww2.amstat.org/sections/srms/Proceedings/y2014/files/311975_88312.pdf.

Witt, M. B. (2010). Estimating the R-indicator, Its Standard Error and Other Related Statistics with SAS and SUDAAN. *Proceedings of Survey Research Methods Section*. Alexandria, VA: The American Statistical Association. Retrieved from http://ww2.amstat.org/sections/srms/Proceedings/y2010/Files/309481_61666.pdf.

Online Appendix B: Benchmark comparison articles in the meta-analysis

Alvarez, R. M., Sherman, R. P., & van Beselaere, C. (2003). Subject Acquisition for Web-Based Surveys. *Political Analysis*, 11(1), 23–43.

Barr, M., van Ritten, J. J., Steel, D. G., & Thackway, S. V. (2012). Inclusion of Mobile Phone Numbers into an Ongoing Population Health Survey in New South Wales, Australia: Design, Methods, Call outcomes, Costs and Sample Representativeness. *BMC Medical Research Methodology*, 12(177), 1–8.

Bandilla, W.; Bosnjak, M., & Altdorfer, P. (2003). A Comparison of Web-Based and Traditional Written Self-Administered Surveys Using the ISSP Environment Module. *Social Science Computer Review*, 21(2), 235–243.

Bethlehem, J. & Scherpenzeel, A. C. (2011). How Representative Are Online Panels? Problems of Coverage and Selection and Possible Solutions. In M. Das, P. Ester, L. Kaczmirek (Eds.) *Social and Behavioural Research and the Internet: Advances in Applied Methods and Research Strategies*. New York: Routledge, 105–129.

Blasius, J. & Brandt, M. (2010). Representativeness in Online Surveys through Stratified Samples. *Bulletin de Méthodologie Sociologique*, 107(1), 5–21.

Chang, L., & Kroshnick, J. A. (2009). National Surveys via RDD Telephone Interviewing Versus the Internet: Comparing Sample Representativeness and Response Quality. *Public Opinion Quarterly*, 73(4), 641–678.

Emerson, M. O., Sikkink, D., & James, A. D. (2010). The Panel Study on American Religion and Ethnicity: Background, Methods, and Selected Results. *Journal for the Scientific Study of Religion*, 49(1), 162–171.

Fitzpatrick, J., Sharp, E. A., & Reifman, A. (2009). Midlife Singles' Willingness to Date Partners with Heterogeneous Characteristics. *Family Relations*, 58(1), 121–133.

- Goldenbeld, C., & de Craen, S. (2013). The Comparison of Road Safety Survey Answers between Web-Panel and Face-to-Face: Dutch Results of SARTRE-4 Survey. *Journal of Safety Research*, 46(1), 13–20.
- Johnson, K. (2005). Effects of Question Order on Estimates of the Prevalence of Drinking and Driving and Seatbelt Use among Region of Peel Rapid Risk Factor Surveillance System Respondents. *Report prepared for the Region of Peel Public Health*. Retrieved from <http://www.rrfss.ca/resources/RRFSS%20Order%20Effects%20Report,%20Peel%20Region,%20Aug%202005.pdf>.
- Keeter, S., Miller, C., Kohut, A., Groves, R. M., & Presser, S. (2000). Consequences of Reducing Nonresponse in a National Telephone Survey. *Public Opinion Quarterly*, 64(1), 125–148.
- Keeter, S., Kennedy, C., Dimock, M., Best, J., & Craighill, P. (2006): Gauging the impact of growing nonresponse on estimates from a national RDD telephone survey. *Public Opinion Quarterly*, 70(5), 759–779.
- Kennedy, C., Mercer, A., Keeter, S., Hatley, N., McGeeney, K., & Gimenez, A. (2016). Evaluating Online Nonprobability Survey. Retrieved from Pew Research Center: <http://www.pewresearch.org/files/2016/04/Nonprobability-report-May-2016-FINAL.pdf>.
- Kohut, A., Keeter, S., Doherty, C., Dimock, M., & Leah, C. (2012). Assessing the Representativeness of Public Opinion Surveys. Retrieved from Pew Research Center: <https://pdfs.semanticscholar.org/b901/168ce90c106703861ce7f8dc2c8eeb4a3458.pdf>.
- Leider, S., & Roth, A. E. (2010). Kidneys for Sale: Who Disapproves, and Why? *American Journal of Transplantation*, 10(1), 1221–1227.
- Linsell, L., Burgess, C. C., & Ramirez, A. J. (2008). Breast cancer awareness among older women. *British Journal of Cancer*, 99(1), 1221–1225.
- Liu, H., Cella, D., Gershon, R., Shen, J., Morales, L. S., & Riley, W. (2010). Representativeness of the PROMIS Internet Panel. *Journal of Clinical Epidemiology*, 63(11), 1169–1178.

Luth, L. (2008, November, 19-21). *An Empirical Approach to Correct Self-Selection Bias of Online Panel Research*. Paper presented at the 5th Annual Data Collection Conference (CASRO), San Diego, California.

Loosveldt, G., & Sonck, N. (2008). An Evaluation of the Weighting Procedures for an Online Access Panel Survey. *Survey Research Methods*, 2(2), 93–105.

Malhotra, N., & Krosnick, J.A. (2007). The Effect of Survey Mode and Sampling on Inferences about Political Attitudes and Behavior: Comparing the 2000 and 2004 ANES to Internet Surveys with Nonprobability Samples. *Political Analysis*, 15(3), 286–323.

Marta-Pedroso, C., Freitas, H., & Domingos, T. (2007). Testing for the survey mode effect on contingent valuation data quality: A case study of web based versus in-person interviews. *Ecological Economics*, 62(1), 388–398.

Martin, S. A., Haren, M. T., Middleton, S. M., & Wittert, G. A. (2007). The Florey Adelaide Male Ageing Study (FAMAS): Design, procedures & participants. *BMC Public Health*, 7(1), 1–15.

Massey, D. S., & Zenteno, R. (2000). A Validation of the Ethnosurvey: The Case of Mexico-U.S. Migration. *The International Migration Review*, 34(3), 766–793.

Potoglou, D., Kanaroglou, P. S., & Robinson, N. (2012). Evidence on the Comparison of Telephone and Internet Surveys for Respondent Recruitment. *The Open Transportation Journal*, 6(1), 11–22.

Potthof, P., Heinemann, L. A. J., & Güther, B. (2004). A household panel as a tool for cost-effective health related population surveys: validity of the "Healthcare Access Panel". *German Medical Science*, 2(1), 1–8.

Purdie, D. M., Boyle, F. M., Dunne, M. P., Cook, M. D., & Najman, J. M. (2002). Health and demographic characteristics of respondents in an Australian national sexuality survey: comparison with population norms. *Journal of Epidemiological Health*, 56(1), 748–753.

- Rao, R. S., Link, M. W., Battaglia, M. P., Frankel, M. R., Giambo, P., & Mokdad, A. H. (2005). Assessing Representativeness in RDD Surveys: Coverage and Nonresponse in the Behavioral Risk Factor Surveillance System. *Proceedings of the Joint Statistical Meeting*, August 7-11, Minneapolis, Minnesota.
- Sallis, J. F., Hofstetter, C. R., Elder, J. P., & Caspersen, C. J. (1989). A Multivariate Study of Determinants of Vigorous Exercise in a Community Sample. *Preventive Medicine*, *18*(1), 20–34.
- Simons, D. J., & Chabris, C. F. (2012). Common (Mis)Beliefs about Memory: A Replication and Comparison of Telephone and Mechanical Turk Survey Methods. *PLOS One*, *7*(12), 1–5.
- Sticht, T., Hofstetter, C. R., & Hofstetter, C. H. (1996). Assessing Adult Literacy by Telephone. *Journal of Literacy Research*, *28*(4), 525–559.
- Watson, N., & Wooden, M. (2013). Adding a Top-Up Sample to the Household, Income and Labour Dynamics in Australia Survey. *The Australian Economic Review*, *46*(4), 489–498.
- Wells, W., Cavanaugh, M. R., Bouffard, J. A., & Nobles, M. R. (2012). Non-Response Bias with a Web-Based Survey of College Students: Differences from a Classroom Survey About Carrying Concealed Handguns. *Journal of Quantitative Criminology*, *28*(1), 455–476.
- Weinberg, J., Freese, J., & McElhattan, D. (2014). Comparing Data Characteristics and Results of an Online Factorial Survey between a Population-Based and a Crowdsourcing-Recruited Sample. *Sociological Science*, *1*(1), 292–310.
- Weyer, S. M., & Werner, J. J. (2010). Characteristics of nurse practitioners interested in participating in a practice-based research network. *American Academy of Nurse Practitioners*, *22*(1), 156–161.
- Wozniak, K. H. (2016). Perceptions of Prison and Punitive Attitudes: A Test of the Penal Escalation Hypothesis. *Criminal Justice Review*, *1*(1), 1–20.

Yeager, D.S., Krosnick, J.A., Chang, L., Lavitz, H.S., Levendusky, M.S., Simpser, A.; & Wang, R. (2011). Comparing the accuracy of RDD telephone surveys and internet surveys conducted with probability and non-probability samples. *Public Opinion Quarterly*, 75(4), 709–747.

Zempel, K. D. (2013). *Putting Police in the box: the effectiveness of data-driven law enforcement* (Master's thesis, University of Wisconsin Oshkosh). Retrieved from: <http://digital.library.wisc.edu/1793/66215>.

Paper 2

**Is it in the method? Testing five measures of survey
representativeness**

Paper 2

Is it in the method? Testing five measures of survey representativeness¹⁵

Abstract

This paper appraises five methods for assessing representativeness in a survey-comparative context using response rates, R-Indicators, Fractions of Missing Information (FMIs), subgroup response rates, and benchmark comparisons. We investigate representativeness comparatively at the example of three probability-based online and mixed-mode panels, which share the same target population, probability sampling methods, face-to-face recruitment and fieldwork agency, but which differ in the details of their implementation, most notably in the way in which the offline population is included in the sample. Within each panel, representativeness is evaluated across two recruitment stages: the face-to-face recruitment stage and the panel registration stage. The panels thus offer a unique quasi-experimental setting to assess different measures of representativeness within and across panels. Ours is the first empirical study to comprehensively assess the added value of the most common measures of representativeness. Our results show high levels of representativeness across the three large-scale probability-based panels and are mostly consistent across measures, except for the FMIs. While all other measures suggest high degrees of representativeness that tend to increase across recruitment steps, FMIs indicate moderate to low degrees of representativeness.

Keywords

Fraction of Missing Information (FMI), panel comparison, probability sample, representativeness, response rates, R-Indicators

¹⁵ This paper is joint work with Tobias Enderle and Annelies Blom. The current version is under review.

Acknowledgements

The authors gratefully acknowledge support from the Collaborative Research Center (SFB) 884 “Political Economy of Reforms” (projects A8 and Z1), funded by the German Research Foundation (DFG) and from the GESIS Panel, funded by the German Federal Ministry of Education and Research. The authors would like to especially thank Jim Lepkowski, Barry Schouten, Joe Sakshaug, and James Wagner for their advice and feedback on early versions of this paper and Julie Korbmacher for her expertise on and help with matching regional auxiliary data.

1. Introduction

Representativeness is the aim of many cross-sectional and panel surveys; however, measuring representativeness comparatively across surveys is difficult. Response rates are frequently used as a convenient measure of representativeness, because of their ease of calculation. However, several studies have shown that the relationship between response rates and representativeness is weak (see e.g., Groves & Peytcheva, 2008). Other indicators of representativeness may be more precise. Benchmark comparisons, for example, compare sample estimates to true population values; subgroup response rates look at response rates by sample subgroups; and Fractions of Missing Information (FMIs) assess the success of imputing values for the nonrespondents. These measures are calculated individually for each benchmark, subgroup, or variable leading to a large number of (possibly contradictory) indicators. R-Indicators were thus developed to provide summary indicators for the representativeness of a sample. However, R-Indicators are complex in their calculation and depend on the availability of informative auxiliary data (see Schouten *et al.*, 2009; Wagner, 2010).

Each of these indicators of representativeness has advantages and disadvantages, and they differ from the other in what they represent and the way they are calculated. In this paper, we appraise five methods for assessing representativeness in a survey-comparative context: response rates, R-Indicators, FMIs, subgroup response rates, and benchmark comparisons. Furthermore, we discuss how differences in results may result from the underlying definitions, assumptions, variables and modelling of the five measures of representativeness rather than from actual differences in survey sample compositions.

We investigate these five representativeness measures at the example of two probability-based online and mixed-mode panels: the German Internet Panel (GIP; www.reforms.uni-mannheim.de/internet_panel) and the GESIS Panel (<http://www.gesis.org/en/services/data-collection/gesis-panel>). When considering online panels, assessments of their representativeness are of importance, because their ability to produce data that reflects the general population of a country has often been questioned (for an overview of the debate see Couper, 2013), in particular where opt-

in panels are concerned (see e.g., Yeager *et al.*, 2011). To meet this criticism, research teams across the globe have set up online panels that are recruited with traditional probability sampling methods and fieldwork techniques and that include the offline population by either equipping them with online devices or by applying mixed-mode designs. Examples of such online panels in Europe are the LISS panel in the Netherlands, the ELIPSS in France and the GIP and GESIS Panel in Germany (see Blom *et al.*, 2016a). Other examples worldwide include the Knowledge Panel in the USA (<http://www.knowledgenetworks.com>) and the new Australian Social Research Centre “Life in Australia” panel (Pennay *et al.*, 2016). The aim of these probability-based online panels is to be representative of the general population in the respective country. The approaches to achieving this goal, however, differ, in particular with regard to the types of sampling frames used and the inclusion of those parts of the population that do not have access to the Internet (see Blom *et al.*, 2016b). Such differences raise questions about which strategies are most effective at reaching representativeness.

The research in this paper originated from a unique position, in which we have access to the survey data, sampling frame data, benchmark information, interviewer observations, and auxiliary geo-coded data of three large-scale probability panel recruitments of the same general population, enabling us to conduct comprehensive comparative analyses into their representativeness. Two of the panels, in fact, share an especially large number of characteristics, because they are two independent samples of the GIP, one recruited in 2012 and the other in 2014. The other panel we examine is the GESIS Panel which was recruited in 2013.

While all three panels (GIP 2012, GESIS Panel, and GIP 2014) are face-to-face recruited probability panels of the general population, including the offline population, they differ in terms of sampling frame used and mode of interview for the offline population, resulting in different recruitment protocols. This enables us to carefully examine the effect of these differences in recruitment protocols on the representativeness of the two panels and to calculate and compare the results of our five indicators of representativeness. In the course of comparatively assessing the

representativeness of the GIP and GESIS Panel, we provide evidence on how insightful and robust the results of these five widely-used measures of representativeness are.

2. Measures of survey representativeness

Two broad perspectives on investigating survey representativeness prevail with regard to theoretical considerations. One is to categorize approaches according to their underlying definition of representativeness. The other way is to structure measures by the level of aggregation that the indicators achieve. Strictly speaking, with regard to defining the concept, survey representativeness is the degree to which the set of respondents to a survey accurately reflects its target population (see left side of Figure 1). Coverage error, sampling error and nonresponse error are all error sources that can cause misrepresentation in a survey response set. In the literature, these error sources are often treated separately, because the underlying mechanisms and measures used to detect these errors may differ across error types (see Biemer, 2010; Groves & Lyberg, 2010; Smith, 2011). How much each individual error source contributes to the overall survey misrepresentation largely depends on the design of a survey. Web surveys, for instance, often have a higher coverage bias than face-to-face surveys, because parts of the survey target population do not use the Internet (see e.g., Couper, 2008; Eckman, 2015).

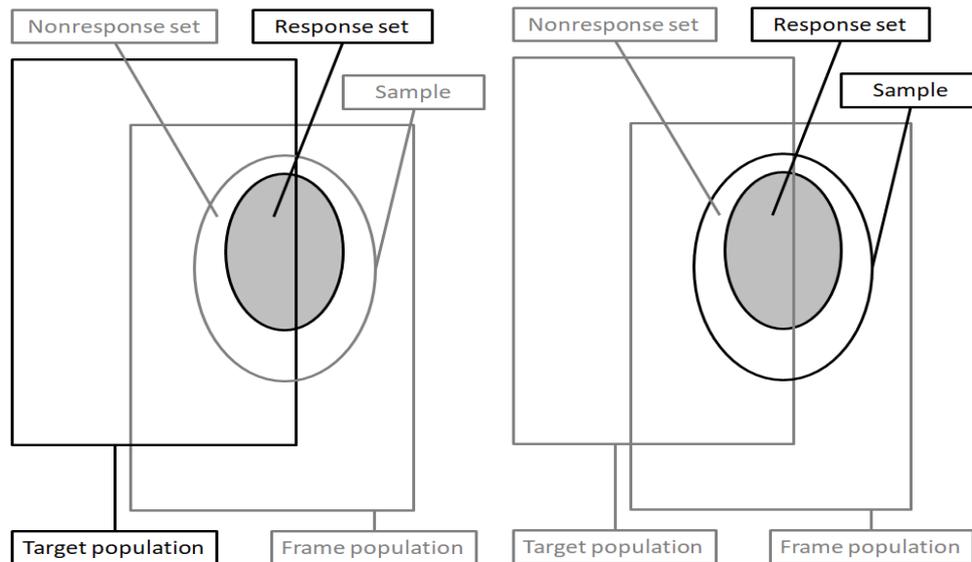


Figure 1: Strict and weak definition of survey representativeness (adapted from Särndal and Lundström, 2005)

Indicators that operationalize the strict definition of representativeness can mostly be found in the survey practice literature, such as field reports and technical surveys reports. The European Social Survey (ESS), for instance, publishes population statistics on its website against which survey representativeness can be assessed (European Social Survey, 2016). Such benchmark comparisons are also used to develop and assess the success of weighting procedures (see e.g., Holt & Smith, 1979; Kalton & Flores-Cervantes, 2003; Kish, 1965; Steinmetz *et al.*, 2014) and to investigate mode effects (see e.g., Malhotra & Krosnick, 2007). Variables that are commonly used for benchmark comparisons are age, gender, education and other key socio-demographic characteristics for which benchmark data are available from official statistics (see e.g., Watson & Wooden, 2013; Fricker *et al.*, 2005; Loosveldt & Sonck, 2008), which are typically very limited and may contain measurement, coverage, sampling and nonresponse errors of their own.

Because of these limitations of benchmark comparisons and because with decreasing response rates (see e.g., de Heer & de Leeuw, 2002; Groves, 2006) nonresponse may be considered the most important source of error leading to misrepresentation, researchers often apply a weaker definition of survey representativeness (see right side of Figure 1). Following this definition, a survey is representative when there is no nonresponse bias. Therefore, it suffices to compare the response set

to the gross sample of a survey, which contains both respondents and nonrespondents. The advantage is that, for surveys with rich sampling frames and linked auxiliary variables, a wider range of indicators can be used to measure survey representativeness. This weaker definition of representativeness is mainly used in the statistical literature on model-based measures for data quality (see e.g., Lundquist & Särndal, 2013; Schouten *et al.*, 2009; Wagner, 2010; Nishimura *et al.*, 2016) and in the typology of nonresponse bias by Wagner (2012).

There are various measures for the weak definition of survey representativeness. The most basic and commonly used measure is the response rate (see AAPOR, 2011). The assumption is that, if the response rate is high, the potential for nonresponse bias is small (see Beullens & Loosveldt, 2010). Indeed, the response rate places an upper bound on the potential nonresponse bias in a survey variable (for a detailed discussion of this argument see Bethlehem *et al.*, 2011: 186). However, although response rates can be valuable in showing the success and efficiency of fieldwork agencies at interviewing the sample members, they are not a reliable measure of survey representativeness. As Groves and Peytcheva (2008) show, surveys with low response rates can have highly representative response sets if nonresponse is a relatively random phenomenon. Conversely, surveys with high response rates can be biased, if the nonrespondents constitute a highly selective subgroup of the sample (see also Bethlehem, 1988; Groves, 2006).

Subgroup response rates follow the same logic as overall response rates, with the exception that they are calculated separately for each category of a variable. They are applied in fieldwork monitoring to intervene if the response set becomes unbalanced due to unequal response probabilities of subgroups. In addition to response rates and subgroup response rates, there are further measures that assess whether a survey response set is representative of the gross survey sample. These are typically based on modelling the difference between respondents and nonrespondents, such as R-Indicators (Schouten *et al.*, 2009; Shlomo *et al.*, 2012) and FMIs (Wagner, 2010).

In this paper, we assess the representativeness of the GIP 2012, GESIS Panel, and GIP 2014 using benchmark comparisons to operationalize the strict definition of representativeness, and response rates, R-Indicators, FMIs, and subgroup response rates to operationalize the weak definition of representativeness. These measures belong to the most widely used indicators of survey representativeness and measure representativeness at different levels of aggregation: aggregated to the survey level or disaggregated at the variable or category level (see Figure 2).

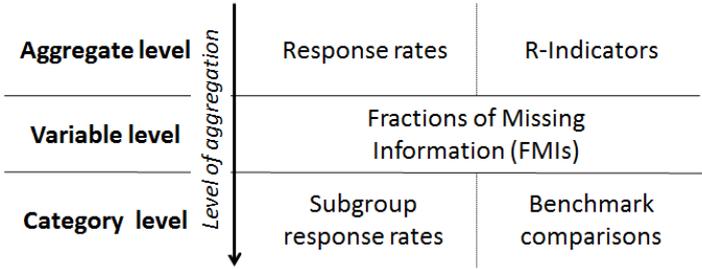


Figure 2: Representativeness measures by level of aggregation

On the aggregate level, response rates and R-Indicators provide a single overall indicator of survey representativeness. This is attractive from a survey documentation point of view, because aggregate indicators can easily be compared across surveys.

Variable level indicators are more detailed and differentiated. Their rationale is that a survey may at the same time accurately reflect the distribution of one variable while it poorly represents another. Depending on the variable of interest, this can mean unbiased estimates for one model and biased estimates for another model. Finally, the most detailed group of indicators investigates representativeness at the category level. Measures at this level are especially valuable during fieldwork where interventions to increase response in an underrepresented subgroup may be possible as well as for post-survey weighting adjustments.

At the aggregate level, we use the two most common measures in our paper: response rates and R-Indicators. At the variable level, we use the highest FMI in a variable. At the category level, we calculate subgroup response rates and conduct benchmark comparisons.

The literature presents innumerable additional measures of representativeness, most importantly Särndal and Lundquist's balance and distance measures (see Särndal & Lundquist, 2014) and partial R-Indicators (see Schouten *et al.*, 2011; Schouten *et al.*, 2012). However, balance and distance measures follow a similar logic as general and partial R-Indicators, and partial R-Indicators are very similar to subgroup response rates (see Shlomo & Schouten, 2013). To limit the conceptual overlap between our measures of representativeness, out of these three measures we chose subgroup response rates as the most commonly used in the literature and survey practice. In the following, each of the aforementioned representativeness measures is discussed in turn.

2.1 Representativeness measures at the aggregate level

2.1.1 Response rates

Response rates are an overall measure of the weak definition of representativeness. They measure the proportion of eligible sample units that was actually interviewed. In general, response rates are thus formalized by

$$RR = \frac{n_R}{n_E} \quad (1)$$

where n_R is the number of respondents and n_E is the number of eligible elements (see Bethlehem *et al.*, 2011).

Response rate are a common measure of representativeness that is easy to compute, interpret and to apply to across surveys. To comparatively assess the representativeness of the GIP 2012, GESIS Panel and GIP 2014, we use AAPOR response rates (see AAPOR, 2011) derived from priority-coded outcomes (see Blom, 2014). The GIP 2012, the GESIS Panel and the GIP 2014 adopt slightly different approaches during the sampling and recruitment processes (see Blom *et al.*, 2016a). Accordingly, the estimation of the reported response rates differs slightly. For the GIP 2012 and GIP 2014 face-to-face recruitment interviews, we use AAPOR response rate 2 (RR2), including short doorstep interviews as partial interviews. With regard to the GIP 2012 and GIP 2014 registration surveys, we report AAPOR

response rate 4 (RR4) assuming 1.78 eligible persons per household in 2012 and 1.77 eligible persons per household in 2014 for households in which the exact number of household members is unknown. All GESIS Panel response rates are based on AAPOR response rate 5 (RR5).

To compare survey representativeness across panels, we provide $100 * (1 - \alpha)\%$ confidence intervals for each estimated measure with

$$CI_{estimate} = estimate \pm z_{\alpha/2} \times SE_{estimate} \quad (2)$$

where $z_{\alpha/2}$ is the $100 * (1 - \alpha/2)$ percentile from the standard normal distribution and $SE_{estimate}$ is the standard error of the estimate.

We compute the standard error of the response rate with

$$SE_{RR} = \sqrt{\frac{RR \times (1 - RR)}{n_E}}. \quad (3)$$

For all reported confidence intervals α is 0.05.

2.1.2 General sample-based R-Indicators

General sample-based R-Indicators are aggregate representativeness measures derived from the predicted response propensities from a logistic regression model of response on a set of auxiliary variables. R-Indicators were proposed by Schouten *et al.* (2009) to compare modified fieldwork strategies experimentally. R-Indicators are used in fieldwork management contexts and general nonresponse bias assessments. Other applications include comparing R-Indicators across experimental conditions in responsive designs (Luiten & Schouten, 2013), comparing R-Indicators over time during the field phase (Beullens & Loosveldt, 2010) and across surveys (Schouten *et al.*, 2012).

R-Indicator values range from 0 to 1. High R-Indicator values mean that response sets are highly representative, whereas low R-Indicator values indicate highly biased response sets. Because of their intuitive value range, R-Indicators are relatively easy to interpret and compare across surveys, time

periods or experimental conditions. However, R-Indicators are strongly influenced by the type, amount, and quality of the auxiliary variables that are fed into the underlying propensity models. In addition, R-Indicators cannot be interpreted as stand-alone values, but meaningful interpretation necessitates the comparison with other R-Indicators (from other surveys, time periods, or experimental conditions), which have to be based on the same propensity model.

Technically, general sample-based R-Indicators are computed as follows. First, the selected auxiliary data are used as independent variables in a logistic response propensity regression model with

$$\text{logit}\left(\frac{\rho_i}{1-\rho_i}\right) = \beta_0 + \beta_1 x_1 + \dots + \beta_k x_k \quad (4)$$

where ρ_i is the response propensity of individual i , β_0 is the model intercept, and β_1 to β_k are the model coefficients for each auxiliary variable x_1 to x_k .

From this model, we predict individual response propensities (see Table A1 in the Appendix). Subsequently, the response propensities are fed into the general R-Indicator formula with

$$R(\rho) = 1 - 2S(\rho) \quad (5)$$

where $S(\rho)$ is the standard deviation of the average response propensity with

$$S(\rho) = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (\rho_i - \bar{\rho})^2} \quad (6)$$

where n is the number of cases, i is the case indicator, ρ is the propensity to respond and $\bar{\rho}$ is the mean response propensity that is roughly equal to the overall survey response rate (RR).

In this paper, we compute R-Indicators for the GIP2012, GESIS Panel and GIP2014 face-to-face recruitment interviews and registration surveys. In our modelling, we take into account the complex survey design of the panels by including the sample stratification characteristics in the logistic regression models underlying the R-Indicator computations. The clustering by sample points, however, cannot be taken into account because including this nominally scaled variable with 250 to

270 categories per model exceeds available computation power. For each R-Indicator, we include a 95% confidence interval based on the bootstrapped standard errors as proposed by Schouten *et al.* (2009) with

$$SE_{R(\rho)} = \sqrt{\frac{1}{B-1} \sum_{b=1}^B (R(\rho)_b - \bar{R}(\rho))^2} \quad (7)$$

where $R(\rho)$ is the R-Indicator, B is the number of bootstrap replications (500 in our analyses), b through B are the bootstrap counters, $R(\rho)_b$ is the R-Indicator for a specific bootstrap and $\bar{R}(\rho)$ is the mean R-Indicator across all B bootstraps.

Since we also compare R-Indicators across the recruitment stages within panels, we have to account for the covariance structure when testing for statistical significance. We therefore take the differences between the bootstrapped R-Indicators of the face-to-face recruitment interviews and registration surveys to receive the mean difference

$$\bar{D}_{R(\rho)} = \frac{1}{B} \sum_{b=1}^B D_{R(\rho),b} \quad (8)$$

where $D_{R(\rho),b}$ is the bootstrap difference $R(\rho)_b^{ecr.int.} - R(\rho)_b^{eg.su}$ in the bootstrap sample of each replication b . To compute empirical confidence intervals, we use the standard error of the difference in R-Indicators across recruitment stages with

$$SE_{D_{R(\rho)}} = \sqrt{\frac{1}{B-1} \sum_{b=1}^B (D_{R(\rho),b} - \bar{D}_{R(\rho)})^2} \quad (9)$$

This approach serves as a hypothesis test of whether the difference between face-to-face recruitment interview and registration survey R-Indicators is significant. If the confidence interval of the bootstrapped difference does not include 0, the difference between face-to-face recruitment interview and registration survey is statistically significant at the 95%-level.

2.2 Representativeness measures at the variable level

2.2.1 Fractions of Missing Information (FMIs)

FMIs capture the uncertainty due to missing data in multiple imputation contexts. In addition, FMIs have been used as an alternative measure for survey representativeness (Wagner, 2010). The general logic of FMIs as representativeness measures is that the nonrespondents are similar to the respondents if the missing values for the nonrespondents can be imputed based on the survey data of the respondents with high degrees of certainty. Low FMIs mean that underlying models fit the data well. High FMIs mean that the nonrespondents systematically differ from the respondents and may therefore cause the data to be biased.

FMIs were first proposed by Dempster *et al.* (1977) for the purpose of missing-data analysis and later applied by Rubin (1987) for multiple imputations quality assessments. Most conventional imputation routines assume the missing data to be missing at random (MAR). When computing FMIs, the missing values of the nonrespondents are multiply imputed using the data available for the respondents. We augment the missing data a pre-specified number of times M following the recommendations in Graham *et al.* (2007) and Bodner (2008). Then, we compute estimates and their standard errors for each augmented data set. Finally, we pool the results from each imputation into a combined multiple imputation estimate and compute its total variance. The combined multiple imputation estimate is the average of the M estimates and the total variance consists of two components: (1) the within-imputation variance, which reflects the conventional sampling error and (2) the between-imputation variance, which captures the additional variance caused by the uncertainty about the imputed values. The more the augmented data and, hence, the estimates per imputation differ from each other, the larger becomes the between-imputation variance.

For each category of a variable, we compute FMIs with

$$\gamma_M = \frac{\nu+1}{\nu+3} \lambda_M + \frac{2}{\nu+3} \quad (10)$$

where ν is an adjusted estimate for the degrees of freedom and λ_M is the proportion of the variation caused by missing data based on the between-imputation variance with

$$\lambda_M = \left(1 + \frac{1}{M}\right) \frac{B_M}{T_M} \quad (11)$$

where M is the number of imputations, B_M is the between-imputation variance and T_M is the total variance.

We report highest FMIs per variable, because they indicate the potential bias on the variable-level and, hence, how many times a variable would have to be imputed to receive an efficient estimate. The smaller the FMIs, the more confidence can be placed in the assumption that the missingness mechanism is MAR. As a benchmark, Nishimura *et al.* (2016) found in a simulation study that the MAR assumption holds for FMIs smaller than the survey nonresponse rate. Similarly, Andridge and Little (2011) find in their proxy pattern-mixture analysis for survey nonresponse that “[i]f the FMI values are all below the nonresponse rate, then we have very strong covariate information, and can be confident that we have good information to correct bias, even if data are MNAR [missing not at random]” (Andridge and Little, 2011: 163/164).

In our analysis, we estimate FMIs for proportions of a set of socio-demographics based on the respondents to the face-to-face recruitment interview and to the registration survey. Within each survey, these variables are observed for respondents only. The missing values for the nonrespondents are imputed multiply (with $M=30$) using chained equation procedures based on multinomial logit models described by van Buuren and Groothuis-Oudshoorn (2011) and on the predictors in the logistic regression model that we specified for the R-Indicator calculations, which includes sample stratification characteristics that describe the complex survey designs of the panels (see Table A1 in the Appendix). To fully account for the complex survey designs, we additionally include the sample points as clustering information in the modelling. Note that to not artificially distort the FMIs, socio-demographic variables are excluded from imputations of other socio-demographic variables.

2.3 Representativeness measures at the category level

2.3.1 Subgroup response rates

Subgroup response rates measure representativeness in terms of nonresponse bias by displaying differences in gross sample subgroups' propensity to respond to a survey. We compute subgroup response rates across panels and recruitment stages for a common set of characteristics. From each subgroup response rate we subtract the overall response rate. As a result, we receive the over- and underrepresentation in each category relative to the overall response rate.

$$D_{subgroup} = \left(\frac{n_{RC}}{n_{EC}}\right) - \left(\frac{n_R}{n_E}\right) = \left(\frac{n_{RC}}{n_{EC}}\right) - RR \quad (12)$$

where n_{RC} is the number of respondents per category, n_{EC} is the number of eligible elements per category, n_R is the number of respondents to the survey, n_E is the number of eligible elements in the sample and RR is the survey response rate.

We compute 95% confidence intervals based on proportions estimates for each subgroup category. If the deviation of a subgroup response rate from the overall response rate is positive, the characteristic is overrepresented in the response set. If the deviation is negative, the characteristic is underrepresented in the response set. The smaller the deviation of each subgroup response rate is from the overall response rate and, consequently, the closer the deviation is to 0, the more representative is the response set with regard to the characteristic.

2.3.2 Benchmark comparisons

For benchmark comparisons, socio-demographic respondent characteristics are compared to a gold standard survey or official data source that describes the distribution of this characteristic in the population. For each category of each variable we compare proportions between the survey and the benchmark. For the proportion of individuals in a category the benchmark comparison takes the following form:

$$D_{benchmark,c} = \left(\frac{n_{Rc}}{n_R}\right) - \left(\frac{N_C}{N}\right) \quad (13)$$

where n_R is the number of respondents, N is the benchmark estimate and C is the category of a variable.

We compute 95% confidence intervals around the deviations of the panel estimates from the benchmark using the standard errors of the proportions estimates. For our analyses, we draw on benchmarks from the German Mikrozensus (Statistisches Bundesamt, 2016), the official statistics on the general population in Germany.

3. Data

3.1 The survey designs of the GIP 2012, the GESIS Panel, and the GIP 2014: differences and similarities

We assess the representativeness of three probability-based panels that are similar with regard to a number of design features (see Blom *et al.*, 2015; <http://www.gesis.org/en/services/data-collection/gesis-panel>; Blom *et al.*, 2016a). In our analyses, we compare the GIP 2012, the GESIS Panel, and the GIP 2014. Furthermore, for each panel, we compare the representativeness at the face-to-face recruitment stage and the online registration stage (see Table 1).

Table 1: Number of cases by panel and recruitment stage

	GIP 2012	GESIS Panel	GIP 2014
Gross sample	4,878 households	19,676 individuals	9,136 households
Face-to-face recruitment stage	2,543 households	7,599 individuals	4,426 households
Panel registration stage	1,603 individuals	4,938 individuals	3,386 individuals

All three panels are ongoing multi-topic panels with a social scientific focus and strive to be representative of the general population in Germany. The panels were recruited during roughly the same time period with the first GIP recruitment in 2012, the GESIS Panel recruitment in 2013 and a second GIP recruitment in 2014. To gain representative survey samples, the GIP 2012, GESIS Panel, and GIP 2014 recruit their respondents via gold-standard survey methods, i.e. is by established multi-

stage probability-based sampling procedures and face-to-face recruitment interviews. All panel sampling designs are set up to be self-weighting, that is with equal probabilities of selection for all sample members. Moreover, the recruitment fieldwork was subcontracted to the same fieldwork agency and even the same team within the survey organization, relying on largely the same interviewers. All panels include in their samples persons that do not have prior access to the Internet. These so-called offliners are recruited into the panels and invited to fill out questionnaires together with the other panelists. All panels conduct web surveys of approximately 20 to 25 minutes length on a bi-monthly basis. Furthermore, the GIP 2012, GESIS Panel, and GIP 2014 respondents receive invitation letters to the first panel survey soon after the face-to-face recruitment interview. This letter contains a web link to the first panel survey. In this first self-administered panel survey, respondents provide further information about their socio-demographic profile. Every individual that completes this last step in the recruitment process is regarded a panel member (see Table 2 for an overview of similarities between the panels).

Table 2: Similarities between the panels

GIP2012, GESIS Panel and GIP2014

Multi-topic panel studies with a social scientific focus
 Target population is the general population of Germany
 Approximately same time period (2012-2014)
 Traditional multi-stage probability sampling
 Face-to-face recruitment interviews and subsequent registration surveys
 Same fieldwork organisation with the same interviewer pool
 Bi-monthly panel survey waves of 20 to 25 minutes length
 Inclusion of individuals without Internet access

The GIP 2012 and GIP 2014 were largely sampled and recruited in the same way. Despite their similarities with the GESIS Panel, however, there are a number of differences in the survey designs (see Table 3). While the GIP 2012 and the GIP 2014 use a three-stage area probability sampling procedure with prior listing of households in each primary sampling unit (PSU) and complete listing of age-eligible individuals in each household, the GESIS Panel draws its sample from locally-based registers held by municipalities in a two-stage sampling process. All panel samples are stratified by the federal state, municipality, and degree of urbanity. In addition, the panels are designed to have

equal probability samples. In practice, however, there might be deviations from the planned sampling design. For example, the GESIS Panel has a small amount of substitutions for municipalities that did not cooperate with the sampling request. Therefore, our expectation is that the GIP 2012 and the GIP 2014 are more representative than the GESIS Panel in the recruitment interview. However, recent research also shows that area probability samples with listings along a random route, for which the GIP 2012 and GIP 2014 are examples, violate the equal probability assumption with repercussion for the sample (see Bauer, 2014 and Bauer, 2016).

As a knock-on effect of the different sampling procedures, the panels apply slightly different fieldwork routines in the recruitment interviews. The GIP 2012 and the GIP 2014 sample all age-eligible household members into the online panel. This means that frequently more than one individual per household is eligible for the online panel. However, to not overburden the household at the panel recruitment, any person at the sampled household above the age of 16 can be interviewed for the initial face-to-face recruitment interview. During this interview, the respondent provides key socio-demographic information on themselves and all other household members. Subsequently, all household members within the eligible age range that consent to be re-contacted are invited to the online panel. In contrast, the GESIS Panel selects age-eligible individuals directly from the register. This individual is interviewed face-to-face during the recruitment interview and subsequently in the online panel. Overall, this leads to a key difference in the task of the face-to-face interviewer across panels: While the GIP 2012 and the GIP 2014 interviewers make contact with the household and interview any household member aged 16 or older, the GESIS Panel interviewers need to interview a named, pre-specified individual. This difference in face-to-face recruitment procedures leads us to the expectation that the GIP 2012 and the GIP 2014 are more representative than the GESIS Panel regarding survey participant household characteristics, because it should be easier to achieve a high response rate and balanced sample of households in the GIP 2012 and the GIP 2014.

Table 3: Main differences between the panels

GIP2012 and GIP2014	GESIS Panel
Area sample of individuals clustered within households and regions	Register-based sample of individuals clustered within regions
Face-to-face recruitment interview with any household member above the age of 16	Face-to-face recruitment interview with named, pre-specified individual
All age-eligible household members invited to the online panel	The pre-specified individual (typically one per household) is invited to the online panel
All panel members interviewed in online mode (offline households are provided with equipment)	Panel members interviewed in mixed-mode (choice of online or paper-and-pencil)

Another major difference between the panels is their approaches to including individuals that do not have access to the Internet. In the GIP, offline households are provided with an Internet connection and the necessary equipment to participate in the online surveys. The GESIS Panel, in contrast, applies a mixed-mode design, where the web is the primary mode, but individuals that cannot or prefer not to fill out the questionnaires online can participate by means of self-administered paper-and-pencil mail questionnaires. Our expectation is that the GESIS Panel is more representative at this second recruitment step, because people who do not have Internet access might either not want to receive Internet access, for example for data protection reasons, or might find it too difficult to answer questions online, because they might not possess the necessary technical skills (Blom *et al.*, 2016b; Leenheer & Scherpenzeel, 2013). We expect that this group of offliners is more likely to participate via mail surveys than to accept and use a technical device with Internet access.

3.2 Auxiliary data for modelling representativeness

To assess the representativeness of the GIP 2012, the GESIS Panel, and the GIP 2014, we use a number of different data sources available for all panels: population statistics from the German Mikrozensus, sampling frame data, interviewer observations and official as well as commercial micro-geographic area data (see Table 4).

Table 4: Auxiliary data by data source and type of analysis

Source	Auxiliary data	Used for
Mikrozensus data	Geographic region, degree of urbanity, age, gender, education, household size	Benchmark comparisons, FMIs
Sampling frame data	Geographic region, degree of urbanity	R-Indicators, subgroup response rates, FMIs
Interviewer observations	Presence of an intercom, type of building, building condition, social class	R-Indicators, subgroup response rates, FMIs
Official micro-geographic area data	Percentage people aged 0-5, percentage immigrants, total balance of people that move to or from the district, aggregated household income	R-Indicators, subgroup response rates, FMIs
Commercial micro-geographic area data	Unemployment rate	R-Indicators, subgroup response rates, FMIs

The German Mikrozensus is a repeated cross-sectional face-to-face survey conducted annually with a probability sample of one percent of the German population. The data from the Mikrozensus are used to derive official population statistics for Germany. Since the Mikrozensus is mandatory, response rates are at least 95% every year (Statistisches Bundesamt, 2016). The core questionnaire contains socio-demographic characteristics. The socio-demographics collected in the GIP 2012, the GESIS Panel, and the GIP 2014 are based on the harmonized questions of the Demographic Standards (“Demographische Standards”) developed by a collaboration of survey methodologists and the German Federal Agency of Statistics (Statistisches Bundesamt, 2010). As a consequence, the Mikrozensus lends itself well as a benchmark to study the representativeness of the three panels, although it should be kept in mind that these benchmark data can also contain survey errors, because the Mikrozensus is a survey itself. In this paper, we use population estimates of the

following socio-demographic characteristics as benchmarks: geographic region, degree of urbanity, age, gender, education and household size.

Furthermore, the GIP 2012, GESIS Panel, and GIP 2014 sampling frames include variables on all of the gross sample members, i.e. means for respondents and nonrespondents. However, since the sampling frames of the panels differ (see above), only the following sampling frame variables were available for all panels: geographic region, sample point, and degree of urbanity. All panel samples are stratified by these variables. By including most, and where possible all, of these variables in our models, we therefore account for the complex survey design of the three panels. The reliability of these data is high and there are no missing values on the variables in any of the panels.

In addition, all three panels instructed their interviewers to record key information about the house and area that the sampled unit lives in. These interviewer observations are available for the gross sample, i.e. means for both respondents and nonrespondents. The variables available are presence of an intercom, type of building, building condition, and predicted social class of the inhabitants. These variables, though insightful and generally meaningful in nonresponse analyses (see Sinibaldi *et al.*, 2014), have to be considered with caution. Some of these assessments are subjective. In addition, these variables contain missing values, ranging between 2 and 20 percent of the cases depending on the variable. We multiply impute the missing data on the interviewer observations (with $M=30$) using a chained equations procedure with predictive mean matching for unordered categorical data (Meinfelder and Schnapp, 2015) for all panels.

Finally, we link information on the micro-geographic area surrounding the sample unit from official and commercial data bases. The German Federal Institute for Research on Building, Urban Affairs and Spatial Development (BBSR) within the German Federal Office for Building and Regional Planning (BBR) provide an online platform called INKAR (Indicators and Maps for City and Spatial Research; Bundesinstitut für Bau-, Stadt- und Raumforschung, 2015) with data on numerous characteristics of the administrative districts. We link INKAR data to our survey data based on the district of the sample

member's address. The indicators we use were originally gathered in the German Census 2011 and updated based on expansion factors provided by the German Federal Statistical Office. These area variables are percentage of people aged 0 to 5, percentage of immigrants, total balance of people that move to or from the district, and aggregated household income. There are no missing values on these variables, but it is possible that there are survey errors in the data.

The most commonly used source for commercial micro-geographic area data in Germany is microm (microm Consumer Marketing, 2013). Microm sells data on different geographic levels mainly for consumer marketing for companies. They provide a large amount of geographic area data of interest to our analyses. However, microm is secretive about how they derive the indicators they provide. Since this secrecy contrasts with transparency standards in academic research, we only use one microm variable in our models, where we could identify and trust the source: the unemployment rate on the district level, which is originally provided by the German Federal Employment Agency (BA). There are no missing values on this variable in the GIP 2012, 5 percent missing values in the GESIS Panel, and 1 percent missing values in the GIP 2014 due to address mismatches. Since missing data rates are low, we conduct complete-case analyses.

In a model validation study, we assess the correlations of each auxiliary variable with survey participation as well as socio-demographic characteristics (age, gender, education, household size). All correlations are relatively weak. The only correlations with a coefficient of more than 0.1 are with the degree of urbanity in the GESIS Panel recruitment interview, with social class in the GIP 2014 recruitment interview, and the GIP 2012 and GIP 2014 registration survey (see the online appendix for all results of the validation study).

4. Results

In the next section, we juxtapose the five representativeness measures described in section 2. For each measure, we estimate the sample representativeness of the GIP 2012, the GESIS Panel, and the GIP 2014. To observe developments in representativeness during each recruitment process, we

further estimate representativeness at the face-to-face recruitment interview and at the registration interview stages for each of the panels. Following the representativeness measure hierarchy, we first present results on the two aggregate level measures, response rates and R-Indicators, next the results for the highest FMIs as a variable level measure and last, the results of our two category level measures, subgroup response rates and benchmark comparisons.

4.1 Response rates

We find that the response rates are average to high in the GIP 2012, the GESIS Panel, and the GIP 2014 compared to other high-quality face-to-face surveys conducted in Germany, which achieve response rates of 30% to a maximum of 50% (see Pfforr *et al.*, 2015; Stoop *et al.*, 2011 for overviews).

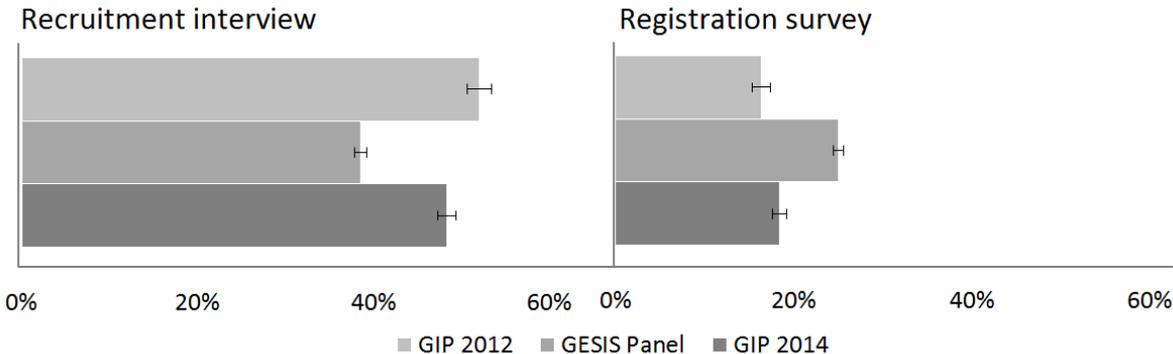


Figure 3: Response rates in the GIP 2012, GESIS Panel and GIP 2014 (GIP 2012 and GIP 2014 recruitment interview: AAPOR RR 2; GIP 2012 and GIP 2014 registration survey: AAPOR RR 4; GESIS Panel: AAPOR RR 5)

Figure 3 shows that at the face-to-face recruitment interview, response rates for the three panels are 52.1% (GIP 2012), 38.6% (GESIS Panel) and 48.4% (GIP 2014). At the registration survey, the cumulative response rates are 16.4% (GIP 2012), 25.1% (GESIS Panel) and 18.5% (GIP 2014).

If we compare the three panels with each other across recruitment steps, we see that the GIP 2012 has the highest response rate at the face-to-face recruitment interview, followed by the GIP 2014 and the GESIS Panel. The GESIS Panel, however, has the highest cumulative response rate at the registration survey, followed by the GIP 2014 and GIP2012. With regard to the response rate as a survey representativeness measure, the results indicate that the GIP 2012 and the GIP 2014 are more

representative at the face-to-face recruitment interview, whereas the GESIS Panel is more representative at the registration survey.

4.2 R-Indicators

R-Indicators across panels and recruitment stages are high. Nonetheless, we find significant differences within as well as across the three samples (see Figure 4). In the upper left part of Figure 4, we display R-Indicators per panel at the face-to-face recruitment interview, in the lower left part we display R-Indicators at the registration survey and the right side shows the bootstrapped differences of R-Indicators across recruitment steps for each panel.

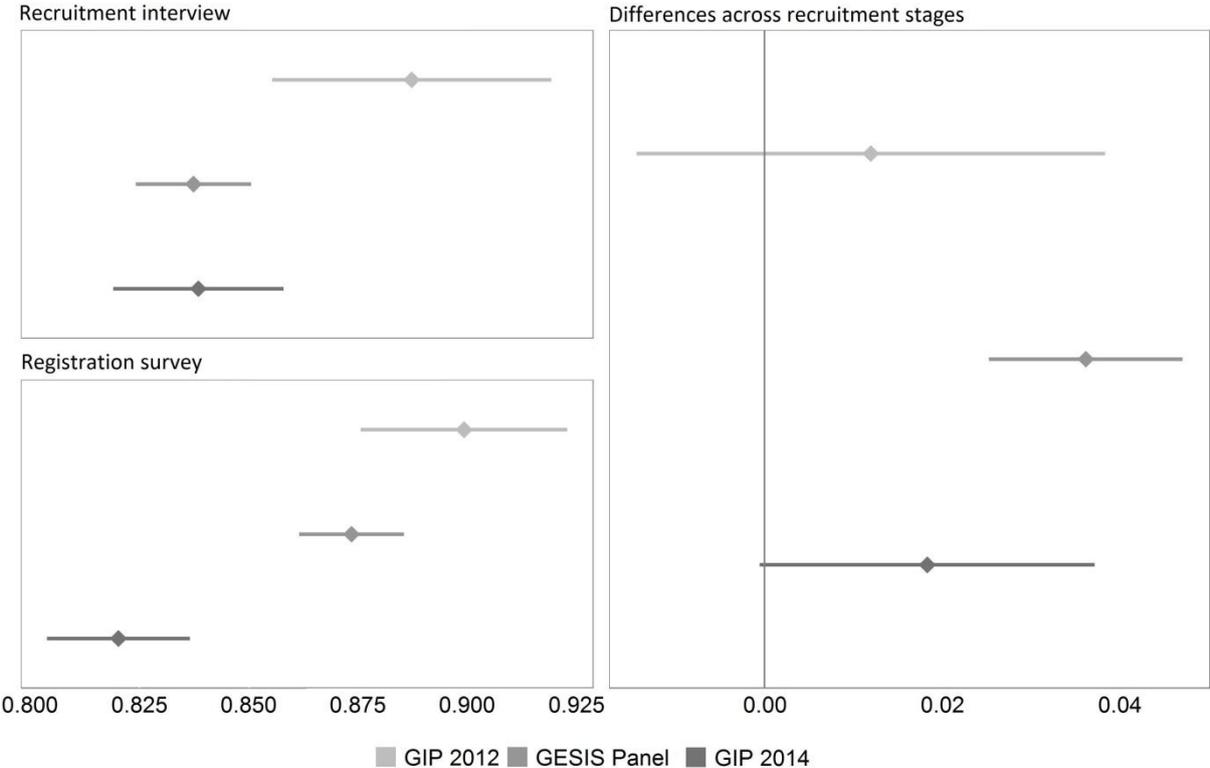


Figure 4: R-Indicators by panel and recruitment stage (including 95% confidence intervals and bootstrapped differences between R-Indicators across recruitment stages; underlying response propensity models contain sampling strata)

At the face-to-face recruitment interview, the R-Indicators for the GIP 2012, GESIS Panel, and GIP 2014 are 0.89, 0.84 and 0.84, respectively. At the registration survey, we find R-Indicators for the GIP 2012, GESIS Panel and GIP 2014 of 0.90, 0.87 and 0.82, respectively. This means that the GESIS Panel significantly increases representativeness from the recruitment to the registration survey (see also

the right pane in Figure 4), while the GIP 2012 and the GIP 2014 remain stable in terms of representativeness across recruitment stages.

When we compare R-Indicators across the three panels and recruitment steps, we see that the GIP 2012 is significantly more representative at the face-to-face recruitment interview than the other two panels. At the registration survey, the representativeness of the GESIS Panel increases significantly and catches up with the GIP 2012. The GIP 2014 shows significantly lower representativeness at the registration survey than the other two panels.

4.3 Fraction of Missing Information (FMI)

We use FMIs as a representativeness measure at the variable level. Figure 5 depicts the highest FMI per socio-demographic variable in the three panels. We use the highest FMI per variable because it tells us whether we can expect the variable to be biased and how many times the variable has to be imputed to get an efficient estimate. Generally, we find that FMIs are moderately high or high (see Figure 5) and are mostly lower than the panel nonresponse rates (see Figure 6). This means that there is considerable misrepresentation on the variables we examine, but for most variables it can be considered as MAR (see Nishimura *et al.*, 2016).

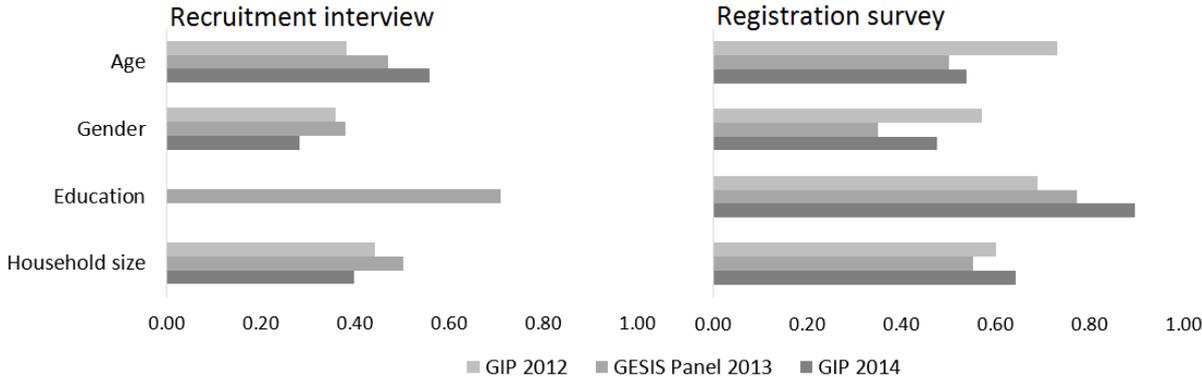


Figure 5: Highest FMIs for each variable by panel and recruitment stage (using 30 imputations with 10 iterations each; imputation models contain sample strata and clustering information; estimates are based on ordered logistic regressions (age, education, household size) and logistic regressions (gender); education not available for GIP 2012 and GIP 2014 recruitment interviews)

Figure 5 shows that, at the face-to-face recruitment interviews, the highest FMIs per variable for the GIP 2012 are 0.38 for age, 0.36 for gender and 0.44 for household size. For the GESIS Panel, they are 0.47 for age, 0.38 for gender, 0.71 for education and 0.50 for household size. For the GIP 2014, the highest FMIs are 0.56 for age, 0.28 for gender and 0.40 for household size.

At the registration survey, the highest FMIs per variable for the GIP 2012 are 0.73 for age, 0.57 for gender, 0.69 for education and 0.60 for household size. For the GESIS Panel, they are 0.50 for age, 0.35 for gender, 0.77 for education and 0.55 for household size. For the GIP 2014, highest FMIs are 0.54 for age, 0.48 for gender, 0.90 for education and 0.64 for household size.

When we compare FMIs across panels and recruitment steps, we find that the GIP 2012 has the lowest FMI for age at the recruitment interview and for education at the registration survey, and the highest FMI for age and gender at the registration survey. The GESIS Panel has the lowest FMI for age, gender, and household size at the registration survey and the highest FMI for gender and household size at the recruitment interview. The GIP 2014 has the lowest FMI for gender and household size at the recruitment interview, and the highest FMI for age at the recruitment interview and for education and household size at the registration survey. The differences across panels, however, are mostly small. From the face-to-face recruitment interview to the registration survey, we find that the FMIs of all panels increase by 0.19 in the GIP 2012, 0.02 in the GESIS Panel and 0.06 in the GIP 2014 on average.

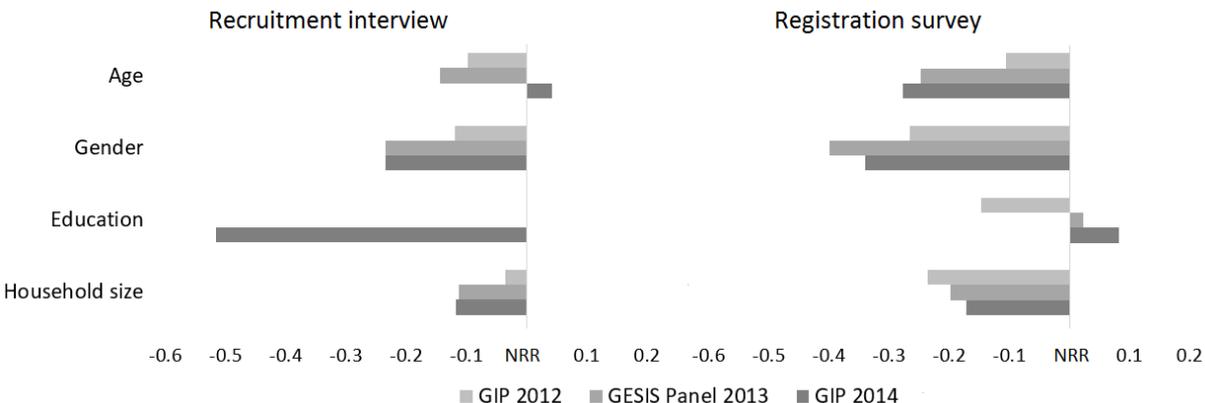


Figure 6: FMI deviations from survey nonresponse rates by panel and recruitment stage (FMIs computed using 30 imputations with 10 iterations each; imputation models contain sample strata and clustering information; estimates are

based on ordered logistic regressions (age, education, household size) or logistic regressions (gender); education not available for GIP 2012 and GIP 2014 recruitment interviews)

Figure 6 depicts the deviations of the largest FMIs per variable from each survey nonresponse rate. We see that most FMIs are below the nonresponse rates. Exceptions are age at the GIP 2014 recruitment interview and education at the GESIS Panel registration survey and the GIP 2014 registration survey. Following the rule of thumb by Nishimura *et al.* (2016), this means that the missingness mechanisms for most but not all of the examined variables can be regarded as MAR across panels and recruitment steps.

4.4 Subgroup response rates

The deviations in subgroup response rates from the overall panel response rates are displayed in Figure 7 (for all subgroup response rates see Table A2 in the Appendix). We compute subgroup response rates for all auxiliary variables. Overall, we find that all panels over- or underrepresent largely the same characteristics. There are few differences across panels and recruitment stages, except for some of the interviewer observations. For the auxiliary variables that we linked from external data sources, representativeness increases from the face-to-face recruitment interview to the registration survey. In the following, we describe the results in more detail. We only report over- and underrepresentations where there is a significant deviation of the subgroup response rate from the overall response rate and where this deviation exceeds three percentage points. If percentage point differences are smaller than three percentage points, we regard the deviation to be of low empirical relevance and thus do not report these in the text, but they are visible in Figure 7.

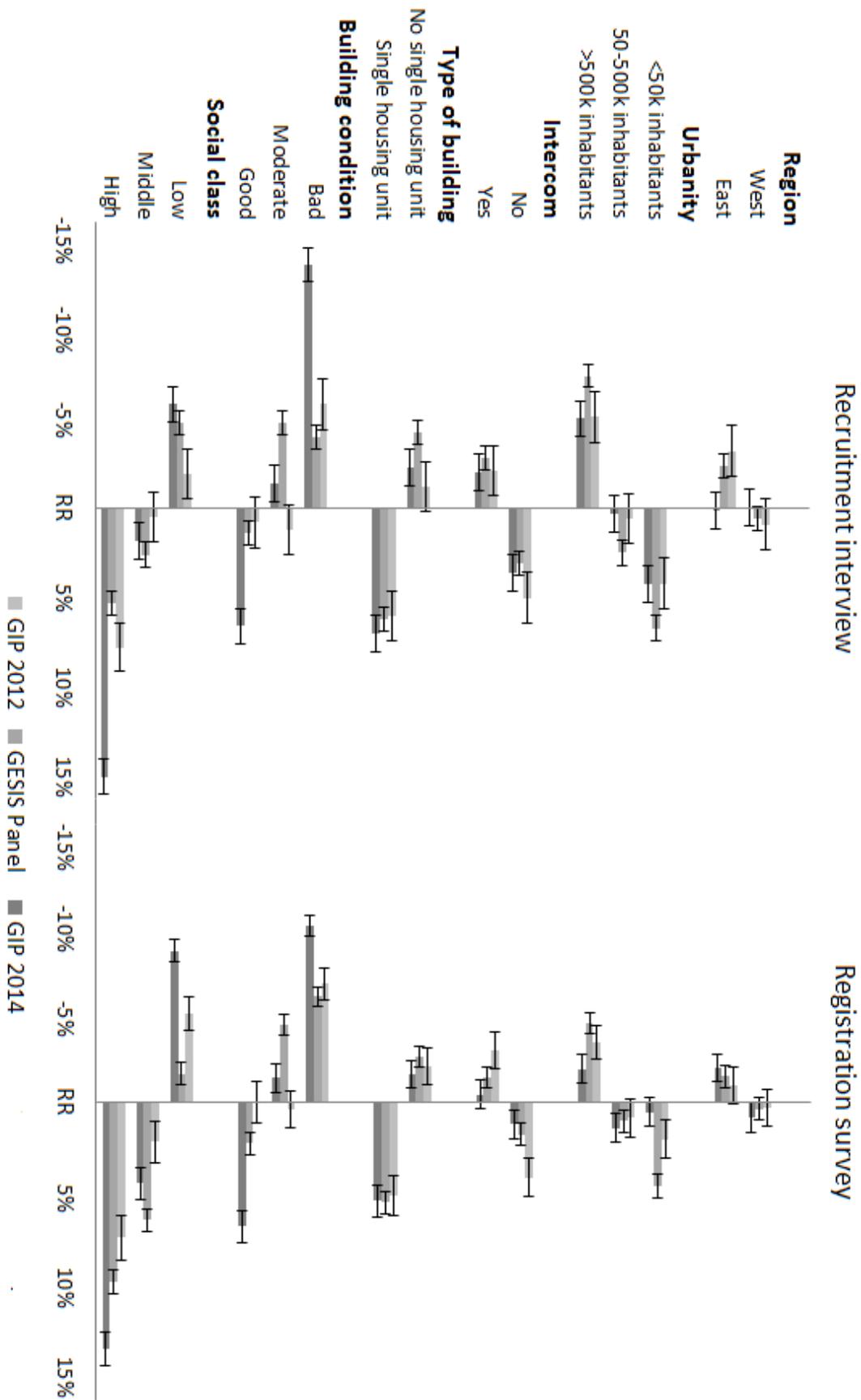


Figure 7: Subgroup response rate deviations from overall response rates by panel and recruitment stage

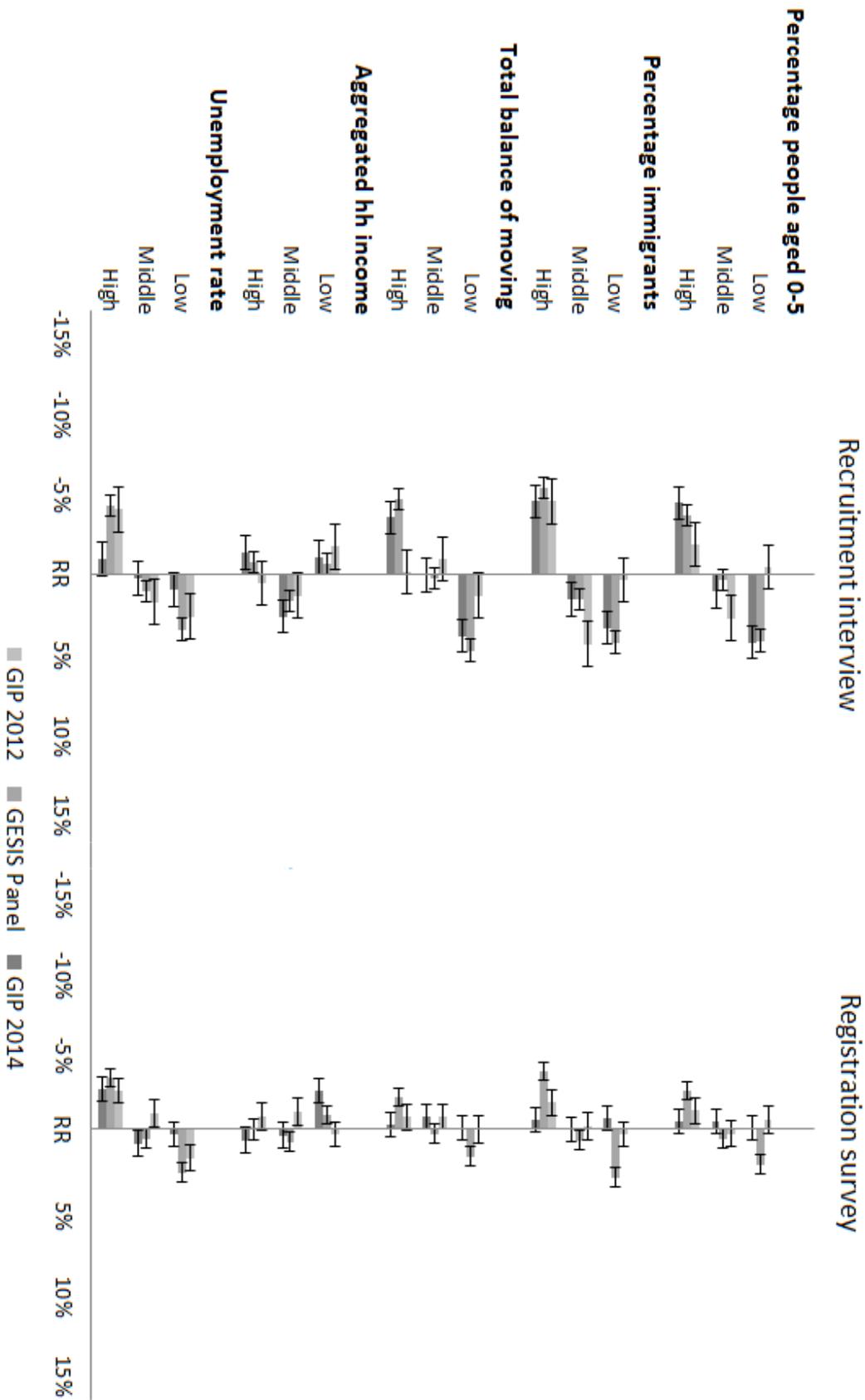


Figure 7 (continued): Subgroup response rate deviations from overall response rates by panel and recruitment stage

At the face-to-face recruitment interview, all three panels show high levels of representativeness for Western and Eastern Germany, medium-sized towns, buildings with an intercom, middle-class individuals, people who live in areas with medium proportions of young children and medium proportions of immigrants, a medium total balance of people moving into and out of the area and areas with a low, medium and high aggregated household income as well as medium unemployment rates (see Figure 7).

All panels significantly overrepresent inhabitants of small towns (GIP 2012 and GIP 2014 by 4%-points, GESIS Panel by 7%-points), single-housing units (GIP 2012 and GESIS Panel by 6%-points and GIP 2014 by 7%-points) and upper-middle-class or higher individuals (GIP 2012 by 8%-points, GESIS Panel by 5%-points and GIP 2014 by 15%-points). Furthermore, the GESIS Panel and GIP 2014 significantly overrepresent people who live in areas with low proportions of young children and low proportions of immigrants (by 4%-points each) and with a low total balance of people moving into and out of the area (by 5%-points and 4%-points).

All three panels significantly underrepresent inhabitants of large cities (GIP 2012 and 2014 by 5%-points, GESIS Panel by 7%-points) and buildings that are in a bad condition (GIP 2012 by 6%-points, GESIS Panel by 4%-points and GIP 2014 by 14%-points). Furthermore, the GESIS Panel underrepresents inhabitants of multi-unit buildings (by 4%-points), buildings that are in a moderate condition (by 5%-points) and lower-class individuals (by 5%-points) and the GIP 2014 overrepresents inhabitants of buildings that are in a good condition (by 7%-points). We also find that all panels underrepresent people who live in areas with high proportions of immigrants (by 5%-points). The GESIS Panel and GIP 2014 underrepresent people who live in areas with high proportions of young children (by 4%-points and 5%-points) and with a high total balance of moving into and out of the area (by 5%-points and 4%-points).

At the registration survey, all panels nicely represent Eastern and Western Germans, inhabitants of medium-sized towns, buildings with an intercom, multi-unit buildings and buildings that are in a

moderate or good condition and all variables that we linked from external data sources. The only minor exception is that the GESIS Panel underrepresents people who live in areas with a high percentage of immigrants (by 4%-points).

All panel samples significantly overrepresent inhabitants of single-housing units (GIP 2012 by 5%-points and GESIS Panel and GIP 2014 by 6%-points), upper-middle-class or higher individuals (GIP 2012 by 8%-points, GESIS Panel by 10%-points and GIP 2014 by 14%-points) and inhabitants of buildings that are in a bad condition (GIP 2012 by 7%-points, GESIS Panel by 6%-points and GIP 2014 by 10%-points). Furthermore, the GIP 2012 and GESIS Panel overrepresent inhabitants of buildings without an intercom (by 5%-points and 4%-points) and the GESIS Panel and GIP 2014 overrepresent middle-class people (by 7%-points and 5%-points). In addition, the GESIS Panel overrepresents inhabitants of small towns (by 5%-points) and the GIP 2012 overrepresents individuals in buildings without an intercom (by 4%-points). The GESIS Panel underrepresents inhabitants of large cities (by 4%-points) and the GIP 2012 and GIP 2014 underrepresent lower-class individuals (by 5%-points and 9%-points).

Overall, across panels, we find a moderate amount of misrepresentation on our observed subgroups and this moderate misrepresentation diminishes over the course of the panel recruitment. The only persistent misrepresentations are found in the degree of urbanity and in some of the interviewer observations (type of building, building condition and social class).

4.5 Benchmark comparison

In Figure 8, we display deviations from the German Mikrozensus of selected socio-demographic characteristics (for all proportions see Table A3 in the Appendix). We find few differences across panels and recruitment steps regarding the degree of representativeness and few variables with a significant misrepresentation. We again report only significant over- and underrepresentations of more than three percentage points to not exceed the scope of the paper.

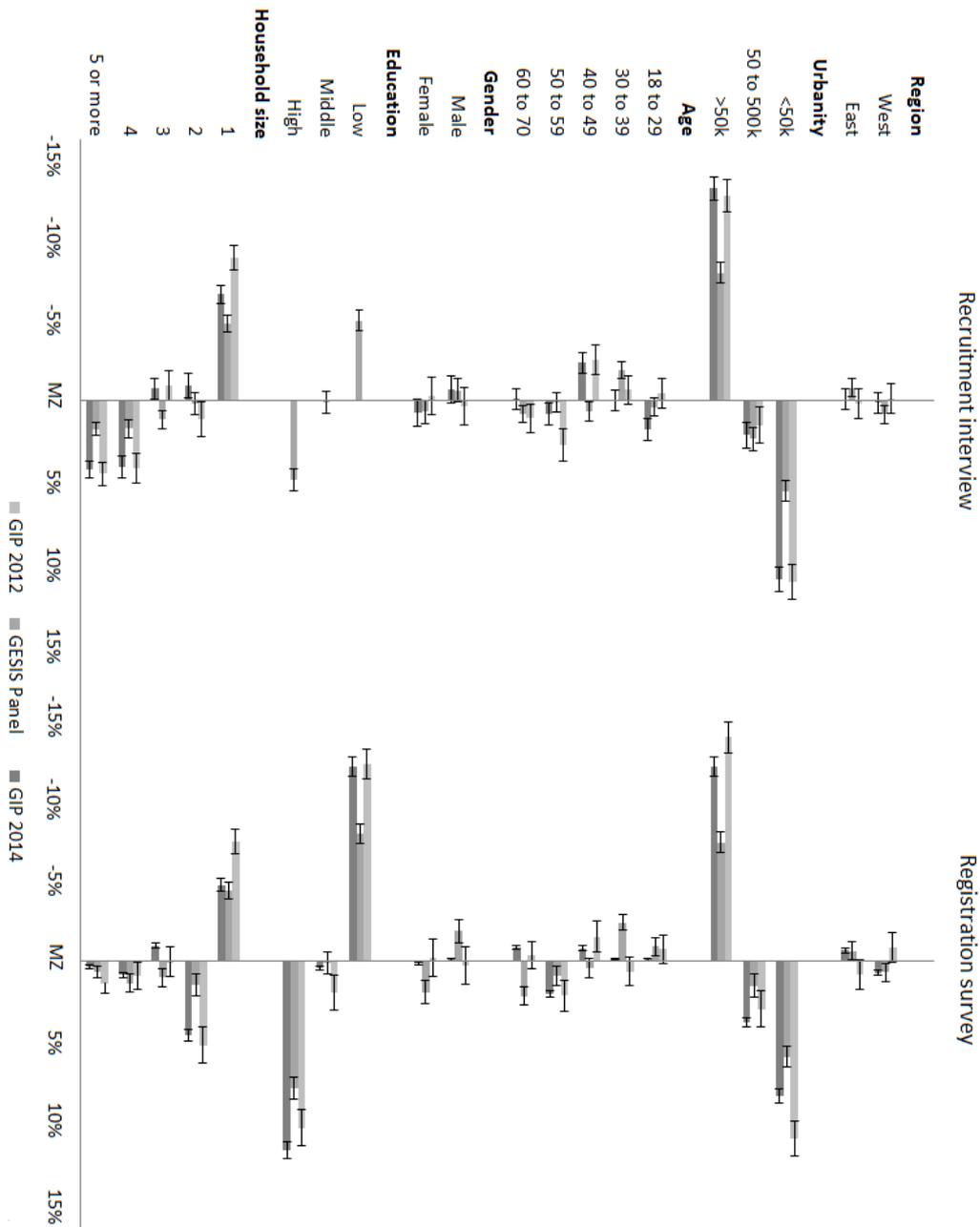


Figure 8: Deviations from the Mikrozensus by panel and recruitment stage (absolute difference between proportions in the survey samples and the Mikrozensus; education not available for GIP 2012 and GIP 2014)

In the face-to-face recruitment interviews, all panels closely represent Western and Eastern Germans, inhabitants of medium-sized towns, all age groups, gender, and households with two or three members. All panels overrepresent small towns (GIP 2012 and 2014 by 11%-points, GESIS Panel by 6%points). They all underrepresent large cities (GIP 2012 and 2014 by 13%-points, GESIS Panel by 6%points). They all underrepresent large cities (GIP 2012 and 2014 by 13%-points, GESIS Panel by 8%-points) and single-person households (GIP 2012 by 9%-points, GESIS Panel by 5%-points and GIP 2014 by 7%-points). The GESIS Panel underrepresents the low educated and overrepresents the high

educated (by 5%- points). There is no data on education for the GIP 2012 and 2014 face-to-face recruitment interviews. The GIP 2012 and 2014 both overrepresent people who live in households with 4 or more members (by 4%-points).

In the registration survey, all panels show high levels of representativeness for Eastern and Western Germans, inhabitants of medium-sized towns, all age groups, gender, people with medium education, and households with three or more members. All panels overrepresent small towns (GIP2012 by 11%-points, GESIS Panel by 6%-points and GIP 2014 by 8%-points) and highly educated persons (GIP 2012 by 10%-points, GESIS Panel by 8%-points and GIP 2014 by 12%-points). They all underrepresent large cities (GIP 2012 by 14%-points, GESIS Panel by 7%-points and GIP 2014 by 12%-points), low educated persons (GIP 2012 and 2014 by 12%-points and GESIS Panel by 8%-points) and single-person households (GIP 2012 by 7%-points, GESIS Panel by 4%-points and GIP 2014 by 5 %-points). In addition, the GIP 2012 and 2014 overrepresent two-person households (by 5%-points).

Overall, across both recruitment steps, we find that all panels represent most of the benchmark characteristics well. The only variables with significant misrepresentation are the degree of urbanity and education. Furthermore, one- and two-person households are slightly misrepresented.

5. Discussion

We divide the discussion into two parts. In the first part, we discuss the representativeness of the three panels relative to each other, the second part deals with the consistency and informative value of the results across the measures tested.

Regarding aggregate level representativeness assessments with R-Indicators, we find that all three panels perform well. The GIP 2012, the GESIS Panel, and the GIP 2014 are all highly representative at the face-to-face recruitment interview (with R-Indicators of 0.89, 0.84 and 0.84, respectively) and, most importantly, at the registration survey (with R-Indicators of 0.90, 0.87 and 0.82, respectively).

The representativeness at the registration stage is especially important, because the response set at this stage is the set of people that is interviewed in future panel waves.

Our results indicate that the GIP 2012 and GIP 2014 samples are more representative at the recruitment interview than the GESIS Panel sample. This is surprising, because recent research shows that random-route procedures as they are used to draw the GIP samples may result in unequal selection probabilities (see Bauer, 2014 and Bauer, 2016), a problem that the register-based sample of the GESIS Panel is not faced with. A potential explanation may be that register-based samples in Germany usually contain substitutions at the level of the municipalities, because some municipalities refuse to draw survey samples. Furthermore, employees at municipalities are typically not sampling statisticians and thus mistakes in drawing the sample may occur. Both aspects of the register sampling process in Germany would explain the lower representativeness of the GESIS Panel at the recruitment interview.

At the registration survey, the GESIS Panel significantly improves its representativeness, while the GIP 2012 and GIP 2014 remain relatively constant. A reason for this could be that the GESIS Panel is more successful at recruiting previously offline sample units into the panel, because of its mixed-mode design. This hypothesis is supported by the fact that in the GESIS Panel 35% of all panel participants use the mail survey mode (see <http://www.gesis.org/en/services/data-collection/gesis-panel>). In contrast, in the GIP 2012 and the GIP 2014, only 8% of the panel participants are previously-offline persons who are provided with the necessary equipment to participate online (see Blom *et al.*, 2016b). Note however, that the GESIS Panel also allows onliners to participate in the paper-and-pencil mode, while the GIP equips only true offliners with devices. This obscures the true difference in offliner recruitment success between the GIP samples and the GESIS Panel.

At the variable level, the FMIs show only small differences in representativeness across the three panels. In addition, the FMIs suggest that the missingness mechanism is mostly MAR (see Nishimura *et al.*, 2016) and can therefore be corrected for. On the category level, the subgroup response rates

as well as the benchmark comparisons find that all three panels overrepresent inhabitants of small towns and single housing units as well as highly educated people and individuals that are rated as upper-middle class or higher by the interviewers. All panels underrepresent inhabitants of large cities and single-person households, as well as people who live in buildings that are rated by the interviewers as being in a bad condition. These findings are consistent with the general literature on survey nonresponse (see e.g., Groves *et al.*, 2002; Bethlehem *et al.*, 2011). Overall, we observe an increase in representativeness from the face-to-face recruitment interview to the registration survey, in particular for the GESIS Panel. This is good news for probability-based online panels and supports the findings by Blom and Herzing (2017).

This paper also assesses the value of five different measures of representativeness. We find that their results are only partially consistent. Response rates and R-Indicators suggest that the GIP 2012 has the highest degree of representativeness at the face-to-face recruitment interview. Furthermore, both response rates and R-Indicators show that the GESIS Panel significantly gains in representativeness at the registration survey stage. This is not surprising, because the mean response propensity, which is approximately equal to the response rate, is part of the R-Indicator formula. Because the literature shows that response rates are poor indicators of nonresponse bias (see Groves, 2006; Groves & Peytcheva, 2008), the R-Indicator should be preferred over the response rate, especially in cases where high-quality sampling frame data is available.

The FMI results, however, are not consistent with these aggregate-level findings. The FMIs suggest only moderate or even low degrees of representativeness across panels. The reason is that the independent variables in our imputation models insufficiently explain the socio-demographic characteristics for which we compute the FMIs. Furthermore, the FMIs suggest a decrease in panel representativeness for all panel samples from the face-to-face recruitment interview to the registration survey. This finding contradicts our aggregate measure findings and can be explained by the logic of the FMIs. Traditionally, FMIs have been used as measures for the efficiency of multiple imputations. Due to respondents dropping out of the study between recruitment steps, the amount

of missing data is higher at the panel registration stage than at the face-to-face recruitment stage. When using the same imputation model on the data for both recruitment steps, the variation of the multiple imputations will therefore be higher at the panel registration stage compared to the face-to-face recruitment stage. For this reason, the increase in the FMIs from the recruitment interview to the registration survey is likely to be an artefact of the method and unrelated to panel representativeness. In addition, FMIs are also highly sensitive to small changes in the modelling. Sensitivity analyses show that small changes in the set of variables used to predict the FMIs result in different findings across surveys that do not follow any systematic pattern. These changes across FMI models are, for example, not in accordance with the correlations between variables that we find in our validation study (see Figures 9 to 11 in the Appendix).

With regard to subgroup response rates, the three panels perform equally well. The results show a general increase in the degree of representativeness from the recruitment interview to the registration survey. The increase in representativeness is especially large for the GESIS Panel. This is consistent with our findings for the R-Indicators. This consistency is highly plausible because both the subgroup response rates and R-Indicators are related to the response rate of a survey and we were able to use the identical set of variables for these analyses.

The benchmark comparisons also show strong similarities in representativeness patterns across the panels. On those categories where we find significant misrepresentation, the GESIS Panel is slightly closer to the benchmark than the GIP samples. Since we largely use the same set of variables for the benchmark comparisons and the FMIs, the results from these two measures should be consistent. The findings are, however, contradictory. Whereas the benchmark comparisons show that all panels represent the age groups and gender well, the FMIs indicate that all panels are only moderately representative for these variables. In addition, the benchmark comparisons suggest that all panels have misrepresentations on education and household size, whereas FMIs indicate that these two variables are better represented in the three panels than age and gender.

Since FMIs are strongly influenced by how well the variables used in the imputation model predict the variables examined, the contradictory FMI results can be attributed to the influences of the imputation model rather than to actual (mis-)representations in the data. FMIs also contradict our findings from the aggregate measures as well as subgroup response rates with regard to the development across recruitment steps. Response rates, R-Indicators and subgroup response rates indicate an increase in representativeness from the recruitment interview to the registration survey. FMIs suggest the opposite, which we attribute to the higher amount of missing data that has to be imputed in the registration survey, rather than to actual differences in response set compositions.

Considering the informative value of the five measures of representativeness examined in this paper, we find that the extent to which they lend themselves to comparative investigations depends on the purpose of the investigation. The advantage of response rates and R-Indicators is that they provide a single aggregate measure, which is especially valuable for identifying general differences between surveys at first glance. An important aspect of response rates is that they do not rely on auxiliary information, which is often unavailable. But response rates are only weakly related to nonresponse bias. R-Indicators are more informative, because they model representativeness using a set of auxiliary variables as predictors. In addition, R-Indicators can use any auxiliary information available for respondents and nonrespondents.

Subgroup response rates and benchmark comparisons draw a more differentiated picture of representativeness than response rates and R-Indicators and are useful if researchers aim to investigate which subgroups are misrepresented in the sample. This large degree of differentiation, however, hampers a general comparative assessment of representativeness. Furthermore, finding identical questions across two or more surveys and thus effectively using the power of these lower level measures is difficult. For instance, including some substantive variables in our analyses would have added value to our investigation. Unfortunately, there were no identically measured survey questions in the GIP 2012, GESIS Panel, and GIP 2014 recruitment and registration stages apart from socio-demographics. Especially benchmark comparisons are typically limited to socio-demographic

characteristics (see e.g., Watson & Wooden, 2013; Fricker *et al.*, 2005; Loosveldt & Sonck, 2008). Benchmark comparisons are, however, a meaningful supplement to any response rate based measure, because they allow us to look at survey representativeness from a different angle, one which is more closely related to nonresponse bias analyses. Finally, FMIs hint at the underlying nonresponse mechanism, a feature which can be valuable for bias corrections. However, the uses of FMIs for analyses of representativeness are limited, because FMIs are unsuited to investigate differences in representativeness across surveys or changes over time. The reason for this is that, if the amount of missing data relative to the non-missing data increases, the FMIs increase too, even if the missingness mechanism is MCAR. Furthermore, FMIs highly depend on how well the variables included in the imputation model predict the variable for which an FMI is computed. Thus, it is difficult to compare the degree of representativeness across variables. This is why we do not consider FMIs fit for the purpose of comparing representativeness across surveys or within surveys over time.

In conclusion, our recommendation for survey comparative representativeness analyses is to apply at least one measure at the aggregate level (response rates or, preferably, R-Indicators) in addition to at least one measure at the variable or category level (for example, subgroup response rates or benchmark comparisons). Finally, in any analyses it is crucial to keep in mind the amount, quality and explanatory power of the variables that are available for the analysis, because they have much influence on the results. When comparing surveys to each other or across time, it is necessary to base the analysis on identical variables, models and measures, to ensure the comparability of results.

Literature

- American Association for Public Opinion Research (2011) *Standard Definitions: Final Dispositions of Case Codes and Outcome Rates for Surveys*. Revised 2011 (7th ed.). (Available from https://www.aapor.org/AAPOR_Main/media/MainSiteFiles/Standard_Definitions_07_08_Final.pdf.)
- Andridge, R. R. and Little, R. J. A. (2011) Proxy Pattern-Mixture Analysis for Survey Nonresponse. *Journal of Official Statistics*, **27**(2), 153-180.
- Bauer, J. J. (2014) Selection Errors of Random Route Samples. *Sociological Methods & Research*, **43**(3), 519–544.
- Bauer, J. J. (2016) Biases in Random Route Surveys. *Journal of Survey Statistics and Methodology*, **4**(2), 263–287.
- Bethlehem, J. (1988) Reduction of nonresponse bias through regression estimation. *Journal of Official Statistics*, **4**, 251–260.
- Bethlehem, J. G., Cobben, F. and Schouten, B. (2011) *Handbook of nonresponse in household surveys*. *Wiley handbooks in survey methodology*. Hoboken, N.J.: Wiley.
- Beullens, K. and Loosveldt, G. (2010) R-indicators and Fieldwork Monitoring. *Presentation at Q2010, Helsinki, Finland*. (Available from <http://www.risq-project.eu/papers/beullens-loosveldt-2010a.pdf>.)
- Biemer, P. P. (2010) Total Survey Error: Design, Implementation, and Evaluation. *Public Opinion Quarterly*, **74**(5), 817-848.
- Blom, A.G. (2014) Setting Priorities: Spurious Differences in Response Rates. *International Journal of Public Opinion Research*, **26**(2), 245-255.
- Blom, A. G., Gathmann, C. and Krieger, U. (2015) Setting Up an Online Panel Representative of the General Population: The German Internet Panel. *Field Methods*, **27**(4), 391–408.

- Blom, A. G., Bosnjak, M., Cornilleau, A., Cousteaux, A.-S., Das, M., Douhou, S. and Krieger, U. (2016a) A Comparison of Four Probability-Based Online and Mixed-Mode Panels in Europe. *Social Science Computer Review*, **34**(1), 8–25.
- Blom, A. G., Herzing, J. M. E., Cornesse, C., Sakshaug, J. W., Krieger, U. and Bossert, D. (2016b) Does the Recruitment of Offline Households Increase the Sample Representativeness of Probability-Based Online Panels?: Evidence from the German Internet Panel. *Social Science Computer Review*.
- Blom, A. G. and Herzing, J. M. E. (2017) Face-to-face Rekrutierung für ein probabilistisches Onlinepanel. Einfluss auf die Repräsentativität. In *Methodische Probleme von Mixed-Mode-Ansätzen in der Umfrageforschung* (eds S. Eifler and F. Faulbaum), pp. 99-118. Wiesbaden: Springer.
- Bodner, T. E. (2008) What Improves with Increased Missing Data Imputations? *Structural Equation Modeling: A Multidisciplinary Journal*, **15**(4), 651–675.
- Bundesinstitut für Bau-, Stadt- und Raumforschung (2015) *INKAR: Indikatoren und Karten zur Raum- und Stadtentwicklung*. © BBSR Bonn 2015. Bonn: Bundesamt für Bauwesen und Raumordnung. (Available from <http://www.inkar.de>.)
- Bundesinstitut für Bevölkerungsforschung (2016) *Zahl der Privathaushalte und durchschnittliche Haushaltsgröße in Deutschland, 1991 bis 2030*. (Available from http://www.bib-demografie.de/DE/ZahlenundFakten/13/Abbildungen/a_13_02_durchschnittl_hhgroesse_d_1991_2030.html.)
- Buuren, S. van and Groothuis-Oudshoorn, K. (2011) mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, **45**(3). (Available from <http://doc.utwente.nl/78938/1/Buuren11mice.pdf>.)
- Couper, M. P. (2008) *Designing Effective Web Surveys*. New York, NY: Cambridge University Press.
- Couper, M. P. (2013) Is the Sky Falling? New Technology, Changing Media, and the Future of Surveys. *Survey Research Methods*, **7**(3), 145–156.

- de Heer, E. and de Leeuw, W. (2002) Trends in Household Survey Nonresponse: A Longitudinal and International Comparison. In *Survey Nonresponse* (eds. R. M. Groves, D. A. Dillman, J. L. Eltinge and R. J. A. Little), pp. 41-54. New York: John Wiley & Sons.
- Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977) Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, **39**(1), 1–38.
- Eckman, S. (2015) Does the Inclusion of Non-Internet Households in a Web Panel Reduce Coverage Bias? *Social Science Computer Review*, **34**(1), 41–58.
- European Social Survey (2016) *ESS7-2014 Documentation Report. Edition 3.0*. Bergen: European Social Survey Data Archive, Norwegian Social Science Data Services for ESS ERIC. (Available from http://www.europeansocialsurvey.org/docs/round7/survey/ESS7_data_documentation_report_e03_0.pdf.)
- Fricker, S., Galesic, M., Tourangeau, R. and Yan, T. (2005) An Experimental Comparison of Web and Telephone Surveys. *Public Opinion Quarterly*, **69**(3), 370–392.
- Graham, J. W., Olchowski, A. E. and Gilreath, T. D. (2007) How Many Imputations are Really Needed? Some Practical Clarifications of Multiple Imputation Theory. *Prevention Science*, **8**(3), 206–213.
- Groves, R. M. (2006) Nonresponse Rates and Nonresponse Bias in Household Surveys. *Public Opinion Quarterly*, **70**(5), 646–675.
- Groves, R. M. and Lyberg, L. (2010) Total Survey Error: Past, Present, and Future. *Public Opinion Quarterly* **74**(5), 849-879.
- Groves, R. M. and Peytcheva, E. (2008) The Impact of Nonresponse Rates on Nonresponse Bias: A Meta-Analysis. *Public Opinion Quarterly*, **72**(2), 167–189.
- Groves, R. M., Dillman, D. A., Eltinge, J. L. and Little, R. J. A. (2002) *Survey nonresponse. Wiley series in probability and statistics*. New York: Wiley.
- Holt, D. and Smith, T. M. F. (1979) Post stratification. *Journal of the Royal Statistical Society*, **1**,33–46.

- Kalton, G. and Flores-Cervantes, I. (2003). Weighting methods. *Journal of Official Statistics*, **19**(2), 79–94.
- Kish, L. (1965) *Survey Sampling*. New York, NY: John Wiley & Sons.
- Leenheer, J. and Scherpenzeel, A. C. (2013) Does It Pay Off to Include Non-Internet Households in an Internet Panel? *International Journal of Internet Science*, **8**(1), 17-29.
- Loosveldt, G. and Sonck, N. (2008) An evaluation of the weighting procedures for an online access panel survey. *Survey Research Methods*, **2**(2), 93–105.
- Luiten, A. and Schouten, B. (2013) Tailored fieldwork design to increase representative household survey response: an experiment in the Survey of Consumer Satisfaction. *Journal of the Royal Statistical Society*, **176**(1), 169–189.
- Lundquist, P. and Särndal, C.-E. (2013) Aspects of Responsive Design with Applications to the Swedish Living Conditions Survey. *Journal of Official Statistics*, **29**(4), 557–582.
- Malhotra, N. and Krosnick, J. A. (2007) The Effect of Survey Mode and Sampling on Inferences about Political Attitudes and Behavior: Comparing the 2000 and 2004 ANES to Internet Surveys with Nonprobability Samples. *Political Analysis*, **15**, 286–323.
- Meinfelder, F. and Schnapp, T. (2015) BaBooN: Bayesian Bootstrap Predictive Mean Matching - Multiple and Single Imputation for Discrete Data. R package version 0.2-0. (Available from <https://CRAN.R-project.org/package=BaBooN>.)
- microm Consumer Marketing (2013) *microm Datenhandbuch*. Nuess, DE: microm Micromarketing-Systeme und Consult GmbH.
- Nishimura, R., Wagner, J. and Elliott, M. R. (2016) Alternative Indicators for the Risk of Non-response Bias: A Simulation Study. *International Statistical Review*, **84**(1), 43-62.
- Pennay, D. W., Neiger, D., Lavrakas, P. J., Borg, K. A., Misson, S. and Honey, N. (2016) *Australian Online Panels Benchmarking Study*. Presented at the World Association for Public Opinion Research (WAPOR) 2016, Austin, Texas.

- Pfarr, K., Blohm, M., Blom, A. G., Erdel, B., Felderer, B., Fräßdorf, M., Hajek, K., Helmschrott, S., Kleinert, C., Koch, A., Krieger, U., Kroh, M., Martin, S., Saßenroth, D., Schmiedeberg, C., Trüdinger, E.-M. and Rammstedt, B. (2015) Are Incentive Effects on Response Rates and Nonresponse Bias in Large-Scale, Face-to-Face Surveys Generalizable to Germany? Evidence from Ten Experiments. *Public Opinion Quarterly*, **79**(3), 740-768.
- Rubin, D. B. (1987) *Multiple Imputation for Nonresponse in Surveys*. New York, NY: John Wiley & Sons.
- Särndal, C.-E. and Lundström, S. (2005) *Estimation in Surveys with Nonresponse*. Wiley Series in Survey Methodology. Hoboken, NJ: John Wiley & Sons.
- Särndal, C.-E. and Lundquist, P. (2014) Accuracy in Estimation with Nonresponse: A Function of Degree of Imbalance and Degree of Explanation. *Journal of Survey Statistics and Methodology*, **2**(4), 361-387.
- Schouten, B., Bethlehem, J., Beullens, K., Kleven, Ø., Loosveldt, G., Luiten, A., Rutar, K., Shlomo, N. and Skinner, C. (2012) Evaluating, Comparing, Monitoring, and Improving Representativeness of Survey Response Through R-Indicators and Partial R-Indicators. *International Statistical Review*, **80**(3), 382–399.
- Schouten, B., Cobben, F. and Bethlehem, J. (2009) Indicators for the Representativeness of Survey Response. *Survey Methodology*, **35**(1), 101–113.
- Schouten, B., Shlomo, N. and Skinner, C. (2011) Indicators for Monitoring and Improving Representativeness of Response. *Journal of Official Statistics*, **27**(2), 1-24.
- Shlomo, N. and Schouten, B. (2013) Theoretical Properties of Partial Indicators for Representative Response. (Available from <http://www.risq-project.eu/papers/shlomo-schouten-2013.pdf>.)
- Shlomo, N., Skinner, C. and Schouten, B. (2012) Estimation of an indicator of the representativeness of survey response. *Journal of Statistical Planning and Inference*, **142**(1), 201–211.

- Sinibaldi, J., Trappmann, M. and Kreuter, F. (2014) Which Is the Better Investment for Nonresponse Adjustment: Purchasing Commercial Auxiliary Data or Collecting Interviewer Observations? *Public Opinion Quarterly*, **78**(2), 440–473.
- Smith, T. W. (2011) Refining the Total Survey Error Perspective. *International Journal of Public Opinion Research*, **23**(4), 464-484.
- Statistisches Bundesamt (2010) Demographische Standards: Ausgabe 2010. *Statistik und Wissenschaft*, **17**.
- Statistisches Bundesamt (2016) *Mikrozensus 2014. Qualitätsbericht*. Statistisches Bundesamt, Wiesbaden. (Available from https://www.destatis.de/DE/Publikationen/Qualitaetsberichte/Bevoelkerung/Mikrozensus2014.pdf?__blob=publicationFile.)
- Steinmetz, S., Bianchi, A., Tijdens, K. and Biffignandi, S. (2014) Improving Web Survey Quality: Potentials and Constraints of Propensity Score Adjustments. In *Wiley Series in Survey Methodology. Online Panel Research. A Data Quality Perspective* (eds. M. Callegaro, R. Baker, J. Bethlehem, A. S. Göritz, J. A. Krosnick and P. J. Lavrakas), pp. 273–298. Chichester, UK: John Wiley & Sons.
- Stoop, I., Billiet, J., Koch, A. and Fitzgerald, R. (2011) *Improving Survey Response: Lessons Learned from the European Social Survey*. Chichester, UK: John Wiley & Sons.
- Wagner, J. (2010) The Fraction of Missing Information as a Tool for Monitoring the Quality of Survey Data. *Public Opinion Quarterly*, **74**(2), 223–243
- Wagner, J. (2012). A Comparison of Alternative Indicators for the Risk of Nonresponse Bias. *Public Opinion Quarterly*, **76**(3), 555-575.
- Watson, N. and Wooden, M. (2013) Adding a Top-Up Sample to the Household, Income and Labour Dynamics in Australia Survey. *Australian Economic Review*, **46**(4), 489–498.

Yeager, D. S., Krosnick, J. A., Chang, L., Davits, H. S., Levendusky, M. S., Simpser, A. and Wang, R.
(2011) Comparing the accuracy of RDD telephone surveys and internet surveys conducted with probability and non-probability samples. *Public Opinion Quarterly*, **75**(4), 709-747.

Online Appendix

Table A1: Logit models across by panel and recruitment stage

	Recruitment interview			Registration survey		
	GIP 2012	GESIS Panel	GIP 2014	GIP 2012	GESIS Panel	GIP 2014
Constant	0.17 (0.17)	-0.24 (0.05)	-0.47 (0.16)	-1.96*** (0.34)	-1.07*** (0.06)	-2.12*** (0.23)
Region						
East	-0.10 (0.01)	0.01 (0.00)	0.14 (0.01)	0.23 (0.02)	0.05 (0.00)	0.14 (0.01)
Urbanity (admin.)						
50,000-500,000 inhabitants	-0.12 (0.01)	-0.03 (0.00)	-0.09 (0.00)	-0.06 (0.01)	-0.08 (0.00)	0.01 (0.00)
>500,000 inhabitants	-0.32* (0.01)	-0.32*** (0.00)	-0.12 (0.01)	-0.40 (0.03)	-0.29*** (0.00)	-0.03 (0.01)
Urbanity (pol.)						
20,000-100,000 inhabitants	-0.15 (0.01)	-0.21*** (0.00)	-0.12 (0.00)	-0.03 (0.01)	-0.14 (0.00)	-0.09 (0.00)
>100.000 inhabitants	-0.27 (0.01)	-0.35*** (0.00)	-0.19 (0.01)	-0.05 (0.03)	-0.16 (0.00)	-0.06 (0.01)
Intercom						
Yes	-0.20* (0.01)	-0.10** (0.00)	-0.14** (0.00)	-0.21 (0.01)	-0.05 (0.00)	0.01 (0.00)
Type of building						
Single housing unit	0.09 (0.01)	0.13*** (0.00)	0.14 (0.00)	-0.14 (0.01)	0.14*** (0.00)	0.09 (0.00)
Building condition						
Moderate	-0.02 (0.02)	-0.06 (0.01)	0.33* (0.01)	0.16 (0.06)	0.05 (0.01)	0.38 (0.03)
Good	-0.13 (0.03)	0.04 (0.01)	0.50*** (0.02)	0.17 (0.07)	0.21 (0.01)	0.43 (0.03)
Social class						
Middle	0.08 (0.01)	0.27*** (0.00)	0.31*** (0.00)	0.64*** (0.01)	0.35*** (0.00)	0.93*** (0.00)
High	0.46** (0.02)	0.39*** (0.00)	0.83*** (0.01)	1.00*** (0.03)	0.49*** (0.00)	1.40*** (0.01)
Unemployment rate	0.00 (0.00)	0.00 (0.00)	0.02** (0.00)	-0.01 (0.00)	0.00 (0.00)	0.00 (0.00)
Percentage people aged 0-5	0.13 (0.00)	-0.01 (0.00)	-0.09 (0.00)	0.14 (0.01)	-0.06 (0.00)	-0.08 (0.00)
Total balance of people that move to or from the district	0.01 (0.00)	-0.01 (0.00)	0.00 (0.00)	0.01 (0.00)	0.01 (0.00)	0.01 (0.00)
Percentage immigrants	0.01 (0.00)	0.01 (0.00)	0.00 (0.00)	0.01 (0.00)	-0.01 (0.00)	0.00 (0.00)
Aggregate household income	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)

† *** p<0.001; ** p<0.01; * p<0.05

††Standard errors in parentheses

Table A2: Subgroup response rates by panel and recruitment stage

	Recruitment interview			Registration survey		
	GIP 2012	GESIS Panel	GIP 2014	GIP 2012	GESIS Panel	GIP 2014
Overall response rate	52.1%	38.6%	48.4%	16.4%	25.1%	18.5%
Region						
West	53.0%	39.2%	48.4%	16.7%	25.5%	19.3%
East	48.9%	36.2%	48.5%	15.4%	23.7%	16.6%
Urbanity (admin.)						
<50,000 inhabitants	56.3%	45.3%	52.6%	18.5%	29.8%	19.1%
50,000-500,000 inhabitants	52.7%	41.1%	48.7%	17.3%	26.2%	19.9%
>500,000 inhabitants	47.0%	31.2%	43.4%	13.1%	20.7%	16.6%
Urbanity (pol.)						
<20,000 inhabitants	56.6%	45.1%	52.8%	19.0%	29.3%	21.0%
20,000-100,000 inhabitants	51.3%	37.6%	48.2%	15.7%	24.1%	17.5%
>100.000 inhabitants	47.5%	30.9%	43.9%	14.0%	20.3%	17.2%
Intercom						
No	57.1%	41.9%	52.0%	18.3%	26.9%	19.7%
Yes	50.0%	36.0%	46.4%	14.1%	23.7%	18.1%
Type of building						
No single housing unit	50.9%	35.4%	46.1%	14.6%	22.6%	17.0%
Single housing unit	58.1%	45.8%	55.4%	19.0%	30.7%	24.0%
Building condition						
Bad	46.3%	35.5%	34.8%	10.3%	19.2%	8.7%
Moderate	53.3%	34.7%	47.0%	15.0%	20.8%	17.2%
Good	52.9%	40.9%	55.0%	20.6%	27.4%	25.4%
Social class						
Low	54.2%	39.3%	42.5%	11.4%	23.5%	10.0%
Middle	56.6%	46.7%	50.1%	18.6%	31.7%	23.0%
High	63.9%	49.4%	63.3%	23.9%	35.1%	32.3%

Table A3: Shares of socio-demographic characteristics across the GIP 2012, GESIS Panel, GIP 2014, and German Mikrozensus

	Recruitment interview			Registration survey			
	GIP 2012	GESIS Panel	GIP 2014	GIP 2012	GESIS Panel	GIP 2014	MZ 2013
Region							
West	80.10	81.04	80.34	79.41	80.90	80.92	80.24
East	19.90	18.96	19.66	20.59	19.10	19.08	19.76
Urbanity (admin.)							
<50,000 inhabitants	34.19	28.64	34.03	33.94	28.94	31.31	23.11
50,000-500,000 inhabitants	41.77	42.61	42.37	43.17	41.76	44.00	40.28
>500,000 inhabitants	24.04	28.75	23.60	22.89	29.30	24.69	36.61
Urbanity (pol.)							
<20,000 inhabitants	45.52	48.11	44.43	45.91	48.10	42.75	40.74
20,000-100,000 inhabitants	26.42	27.35	25.72	25.64	27.06	24.98	27.23
>100,000 inhabitants	28.06	24.54	29.86	28.45	24.85	32.27	32.03
Age							
18-29	20.06	20.92	22.26	19.81	19.64	20.42	20.55
30-39	16.81	15.62	17.45	18.13	15.11	17.36	17.51
40-49	20.32	23.53	20.53	21.36	23.24	22.05	22.87
50-59	24.42	21.81	22.56	23.77	22.59	23.72	21.74
60-70	18.38	18.12	17.20	16.92	19.42	16.45	17.32
Gender							
Male	50.27	49.3	49.26	50.16	48.08	49.84	49.96
Female	49.73	50.7	50.74	49.84	51.92	50.16	50.04
Education							
Low	–†	27.46	–†	20.29	24.57	20.45	32.38
Middle	–†	34.42	–†	36.26	34.38	34.71	34.32
High	–†	38.12	–†	43.45	41.05	44.83	33.30
Household size							
1	11.67	15.75	13.91	13.13	16.10	15.76	20.46
2	36.70	35.75	34.63	40.71	36.99	40.09	35.56
3	19.51	21.57	19.68	20.48	21.41	19.46	20.43
4	20.80	18.36	20.71	17.50	17.99	17.51	16.68
5 or more	11.32	8.58	11.07	8.18	7.51	7.18	6.86

†Note: Education not available for GIP 2012 and GIP 2014 recruitment interview.

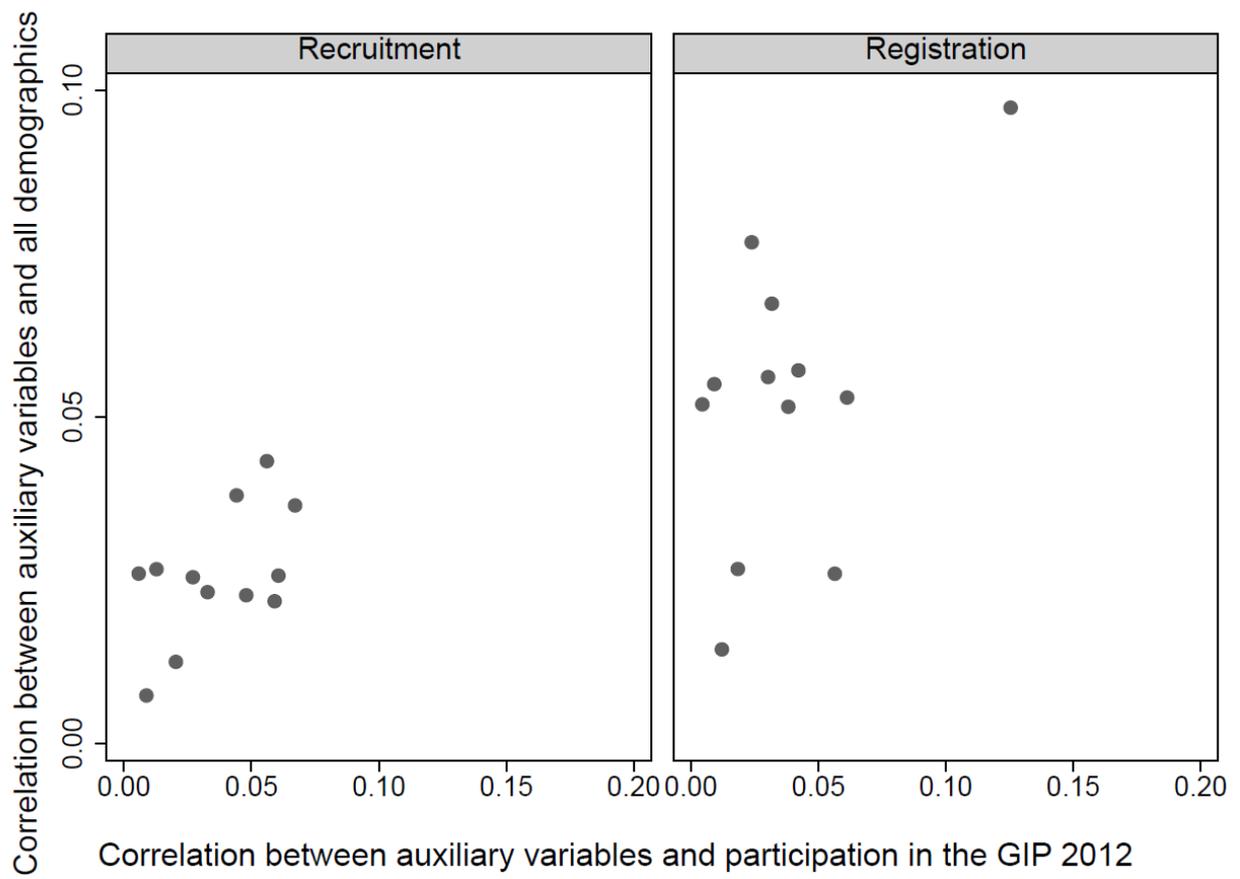


Figure 9: Correlations of panel participation and socio-demographic characteristics with auxiliary variables in the GIP 2012 (demographic variables include age, gender, education and household size; education not available for the recruitment interview)

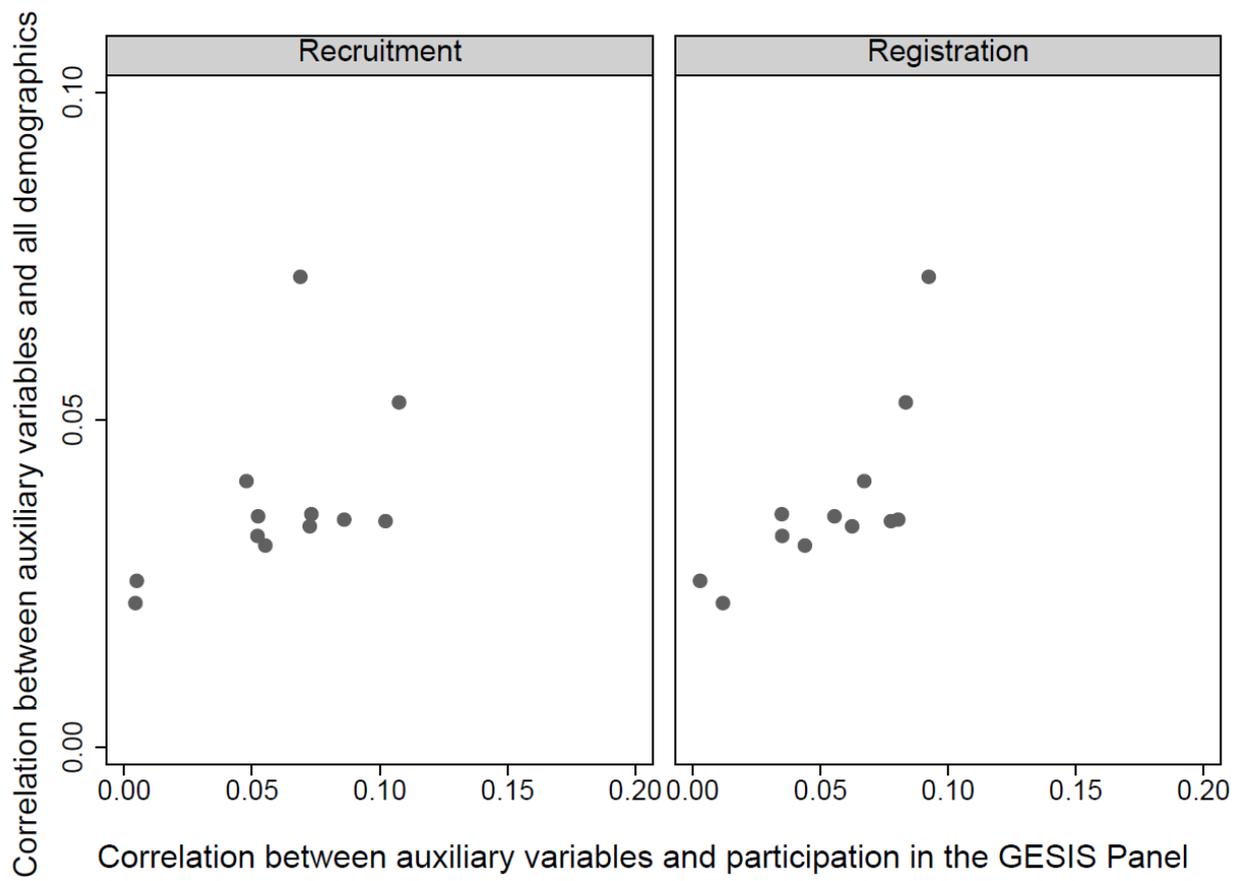


Figure 10: Correlations of panel participation and socio-demographic characteristics with auxiliary variables in the GESIS Panel (demographic variables include age, gender, education and household size)

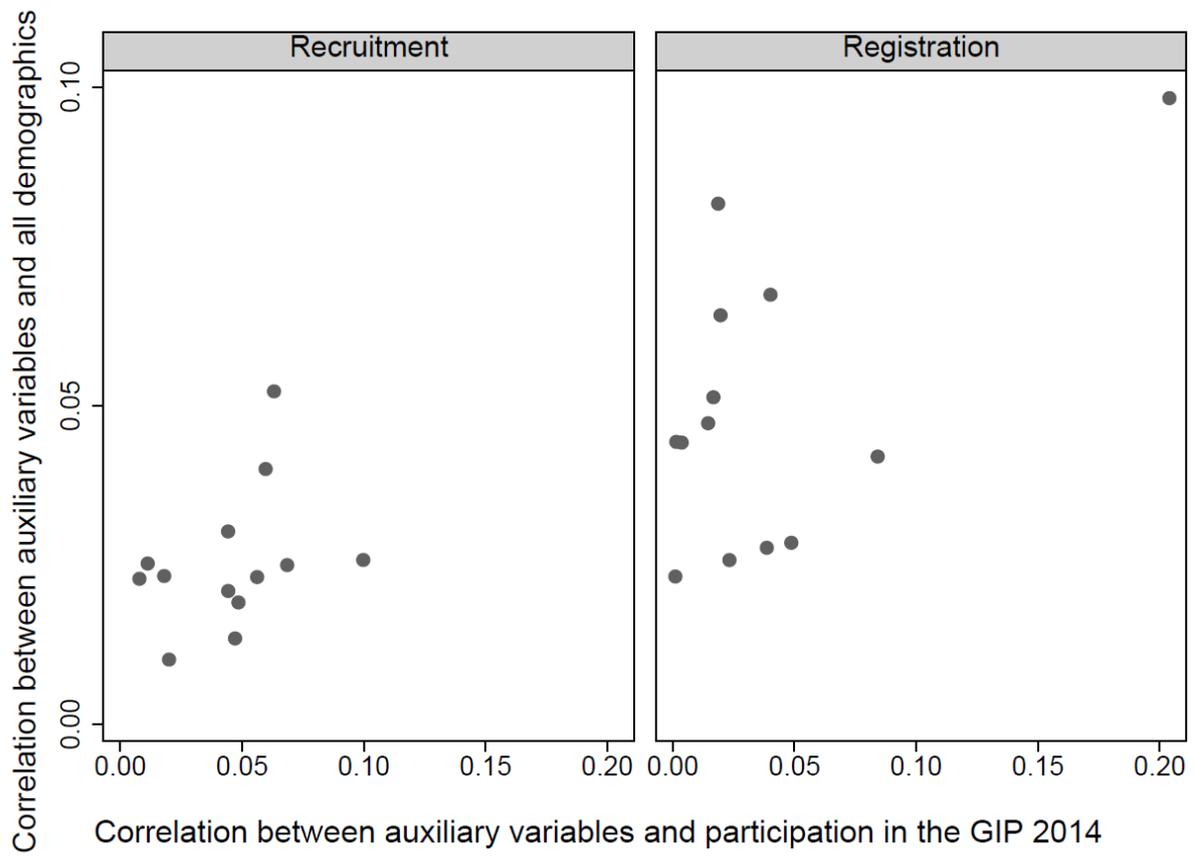


Figure 11: Correlations of panel participation and socio-demographic characteristics with auxiliary variables in the GIP 2014 (demographic variables include age, gender, education and household size; education not available for the recruitment interview)

Paper 3

The utility of auxiliary data for survey response modeling: Evidence from the German Internet Panel

Paper 3

The utility of auxiliary data for survey response modeling: Evidence from the German Internet Panel¹⁶

Abstract

Auxiliary data are becoming more important as nonresponse rates increase and new fieldwork monitoring and respondent targeting strategies develop. In many cases, these auxiliary data are collected or linked to the gross sample to predict survey response. If the auxiliary data have high predictive power, these response models can meaningfully inform survey operations as well as post-survey adjustment procedures. In this paper, I examine the utility of different sources of auxiliary data (sampling frame data, interviewer observations, and micro-geographic area data) for modeling survey response in a probability-based online panel in Germany. I find that each of these data sources are challenged by a number of concerns (scarcity, missing data, transparency issues, and high levels of aggregation) and none of them predict survey response to any substantial degree. Therefore, I conclude that the available auxiliary data should be used with caution. Furthermore, I hope that this paper inspires the search for more predictive auxiliary variables as well as efforts to raise the quality of auxiliary data.

Keywords

Auxiliary data, nonresponse, representativeness, online panel

Acknowledgements

The author gratefully acknowledges support from the Collaborative Research Center (SFB) 884 “Political Economy of Reforms” (projects A8 and Z1), funded by the German Research Foundation (DFG) and from the GESIS – Leibniz Institute for the Social Sciences. I would especially like to thank Annelies Blom for advice and feedback on early versions of this paper.

¹⁶ This paper is single-authored work by Carina Cornesse.

1. Introduction

In light of decreasing response rates and rising concerns about nonresponse bias and cost efficiency, auxiliary data have become popular in survey methodological research in recent years. Auxiliary data are typically external to the survey data collected and for most operational purposes they need to be available for respondents as well as nonrespondents (see for example Sinibaldi et al., 2014). They can be used for a number of survey operational tasks, such as eligibility screening and fieldwork monitoring, as well as post-survey nonresponse adjustments, such as weighting and imputations.

Because of the current high demand for auxiliary data, there is progress in collecting more such data and making them available to secondary data users (for example in the European Social Survey (see ESS, 2012) and in the Health and Retirement Study (see HRS, 2014)). In addition, some official institutions as well as commercial vendors provide aggregated auxiliary data that can be linked to survey data (see Section 3 for more information).

Most approaches to using auxiliary data for nonresponse research rely on the assumption that the data are of high quality (i.e. error-free) and predictive of survey response (as well as, for many purposes, key survey variables; see Olson, 2013 for a formalization of characteristics of auxiliary variables that are useful for nonresponse adjustments). This stands in contrast with the literature on auxiliary data that indicates potential errors in the auxiliary data as well as low predictive power.

This paper contributes to the survey methodological literature on auxiliary data by exploring the utility of different sources of auxiliary data in the context of the recruitment of a probability-based online panel in Germany. I provide an overview of the existing literature on auxiliary data and discuss the advantages and disadvantages of the data sources that were available for my study. Then, I explore whether auxiliary variables are systematically missing by survey response. In addition, I examine the correlations of each of the auxiliary variables with survey response and I assess the predictive power of the auxiliary data as well as the significance of their coefficients in logistic regression models on survey response. I thereby contribute to the scarce literature on assessing the

utility of different types of auxiliary data for predicting survey response. Finally, this study is the first that assesses the utility of auxiliary data for predicting survey response in the context of the different stages of recruiting a panel.

2. Uses and usefulness of auxiliary data

Auxiliary data are used for multiple purposes, many of which involve predicting survey response. One application that has been developed in recent years is model-based representativeness measures such as R-Indicators (see Schouten et al., 2009), balance indicators (see Särndal, 2011), or the Fraction of Missing Information (see Wagner, 2010). These representativeness measures can be reported and compared across surveys or experimental fieldwork conditions after data collection to provide data users with background information on survey data quality (see for example Schouten et al., 2012).

In addition to measuring representativeness after data collection, auxiliary data can be used for fieldwork monitoring during data collection. The National Survey of Family Growth, for instance, has integrated auxiliary data into their fieldwork monitoring “dashboard” (see Kirgis and Lepkowski, 2010). This allows them to detect potential problems (such as high numbers of locked buildings in certain areas) on a daily basis and to then intervene quickly if necessary. Auxiliary data can also be used to monitor whether the sample representativeness is compromised during fieldwork, so that steps can be taken to rebalance the sample, for example by case-prioritization, before the fieldwork phase ends (see for example Schouten, Shlomo, and Skinner, 2011).

A related form of auxiliary data usage is in responsive or adaptive survey designs (see Groves & Heeringa, 2006; Wagner, 2008). The idea of these approaches is that subgroups of the gross sample receive different survey designs depending – at least to some extent – on their response propensity as modeled using auxiliary data. Usually, the most important goal is to reduce variation in response propensities across subgroups of the gross sample. There are many survey design features that can be varied in such a design, for example the survey mode or the timing of contact attempts (see

Schouten et al., 2013 for an extensive list of design features that can be varied in responsive design settings).

In a large-scale study on multiple surveys Schouten et al. (2016) find that these responsive design approaches can reduce nonresponse bias as measured by model-based representativeness measures. Similarly, Wagner et al. (2012) find that responsive design approaches based on the auxiliary data from the NSFG successfully increase response rates among the formerly underrepresented subgroups of the sample leading, as intended, to lower variation in response propensities across subgroups.

Another prominent application of auxiliary data in the literature is nonresponse adjustment weights (see for example Olson, 2013), when cases are weighted by their inverse response propensity calculated from a logistic regression model. Olson (2013) defines a useful auxiliary variable for nonresponse weighting as a variable that is associated with survey response as well as substantive survey variables of interest. In a study on the added value of auxiliary data for nonresponse adjustment weights Kreuter et al. (2010) find that the available auxiliary data were only moderately associated with substantive survey variables and weakly associated with survey response. Nonetheless, the authors report that using the auxiliary data to derive adjustment weights still led to substantial changes in survey estimates. However, not all types of auxiliary data seem to be equally suited to derive nonresponse adjustment weights. Sinibaldi et al. (2013), for example, show that interviewer observations have higher associations than commercial micro-geographic area data with some primary survey variables (unemployment of a household member and income). The interviewer observations therefore contributed more to the nonresponse adjustment weights.

An alternative to weighting procedures is the application of imputation procedures to adjust for survey nonresponse. In this context, West and Little (2013) find that multiple imputations using pattern-mixture models can reduce nonresponse bias even under a potentially non-ignorable missing data mechanism and when the auxiliary data used in the imputation model are error prone.

3. Types of auxiliary data and their quality

There are several types of auxiliary data that are commonly used in the literature and in practice. The most common types are sampling frame data, interviewer observations, and linked micro-geographic area data. In the following, I describe these types of auxiliary data and discuss findings from the literature about the quality of these data.

Sampling frame data are typically available in probability-based surveys, where they describe all units in the gross survey sample. In theory, sampling frame data exist for all units on a frame. However, in practice, researchers only have access to information on the units that were actually drawn into the gross sample. Sometimes sampling frames contain a lot of information, for example when the sample is drawn from a register. In some countries, population registers may include detailed individual information, like age, gender, ethnic background, household composition, employment history, education history, and income. In other situations, however, the sampling frame data may be limited to broad regional information only. In the literature, sampling frame data are commonly used for measuring representativeness (see for example Schouten et al., 2009) as well as fieldwork monitoring (see for example Schouten, Shlomo, and Skinner, 2011) and nonresponse adjustment (see for example Kreuter et al., 2010). However, there is very little research on the quality of sampling frame data (for a discussion of potential data quality issues of sampling frames see for example Hall, 2008).

Interviewer observations contain information about all sample units in the gross sample. This information is recorded by interviewers during fieldwork (see Olson, 2013). Commonly collected interviewer observations include reports about access impediments to the sample unit's house (e.g. closed gates) and the type of housing unit (e.g. single house, terraced house, apartment block, farm, etc.) that the sample unit lives in (see for example West & Kreuter, 2013). Depending on the survey, collected interviewer observations may also include interviewer assessments about the volume of traffic and public transportation stops near the housing unit (like in the Health and Retirement Survey (HRS); see HRS, 2014) or interviewers' guesses about the presence of children in the sampled

household or whether the sample units are sexually active (like in the National Survey of Family Growth (NSFG); see West, 2010).

Some publications consider the quality of interviewer observations. One finding from this field of research is that missing data rates can vary greatly across surveys and are often rather high. Matsuo et al. 2010, for example find that in the European Social Survey the missing data rates for the interviewer observations range from 0.00% in Russia to 40.01% in Germany. Other researchers find that interviewer observations are prone to interviewer effects (see Olson, 2013), especially when the measures collected leave room for interpretation because interviewers can perceive the objects that they observe differently (see for example Raudenbush & Sampson, 1999). West and Kreuter (2013), for example, show that in the NSFG the accuracy of the interviewer observations about the presence of children in the household varies by interviewer experience.

The contribution of interviewer observations to nonresponse adjustments is yet unclear. West, Kreuter, and Trappmann (2014), for example, find that using interviewer observations in nonresponse adjustment weights hardly has an impact substantive survey estimates. The authors suspect that this is due to the low predictive power of the interviewer observations in modeling survey response. However, Sinibaldi et al. (2014) find that interviewer observations were more successful than linked commercial micro-geographic area data in predicting primary substantive survey outcomes and were therefore better suited to inform nonresponse adjustment weighting.

Micro-geographic area data are linked to the survey data from external (official or commercial) sources and contain aggregated information that describe all sample units' environment. Typical data from such sources include aggregate measures of income or purchasing power in each area, household composition in terms of socio-demographics, and ethnic or religious composition of the neighborhood (see for example West et al., 2015). They can often be purchased from commercial marketing vendors, such as Microm¹⁷ in Germany, Experian¹⁸ and Callcredit¹⁹ in the UK, and

¹⁷ www.microm.de

MSG²⁰ and Aristotle International Inc.²¹ in the US. There are, however, also official institutions that make micro-geographic area data available, such as the German Federal Institute for Research on Building, Urban Affairs and Spatial Development (BBSR) in Germany, the Office for National Statistics (ONS) in the UK, and the United States Census Bureau in the US.

Micro-geographic area data are most commonly used in nonresponse adjustment weighting. The literature, however, shows that micro-geographic area data usually correlate little with survey response (see for example West et al., 2015) and are therefore not useful in nonresponse adjustment procedures (see for example Biemer & Peytchev, 2013). However, there are also studies that show that including micro-geographic area data in nonresponse adjustment weights can lead to shifts in survey estimates, especially when there is at least a moderate association with substantive survey variables (see for example Kreuter et al., 2010).

A potential disadvantage of the commercial micro-geographic area data is that they have been found to be prone to errors. In a study on the quality of commercial marketing data, Pasek et al. (2014), for example, find that these data were often inaccurate and systematically incomplete. In addition, the authors point to transparency problems with commercial data: “Because these data are of considerable value to the private companies that aggregate them [...] social scientists seem unlikely to gain a full picture” (p. 912). In addition, West et al. (2015) find only weak agreement between identical variables in the data purchased from two different commercial data vendors. However, they also show that the commercial data can effectively be used to predict sample eligibility as well as some substantive survey variables. They, therefore, conclude that buying micro-geographic area data might be a good investment for some survey operational tasks, such as eligibility screening, but less so when it comes to nonresponse weighting.

¹⁸ www.experian.co.uk

¹⁹ www.callcredit.co.uk

²⁰ www.m-s-g.com

²¹ www.aristotle.com

4. Data and methods

In this section, I describe the data that I use for the analyses, the auxiliary data that I use as predictors in the response models, and the methods I apply in the analyses.

4.1 The German Internet Panel (GIP)

For the analyses, I use data from the GIP recruitment phase. The GIP is a probability-based online panel of the general population with bi-monthly panel waves on multiple topics in the social sciences and economics (see Blom, Gathmann, Krieger, 2015). It is based on a three-stage stratified probability area sample, for which areas in Germany are randomly sampled, subsequently all addresses are listed within sampled areas, and then a random selection of households is made from the addresses listed.

The GIP recruitment was conducted in two phases: a face-to-face recruitment interview phase and a subsequent online profile interview (enrollment in the online panel). All age-eligible household members in a household that participated in the recruitment interview were invited to participate in the subsequent online panel waves. People who did not have access to the Internet or Internet-enabled devices were provided with the necessary equipment (see Blom et al., 2016). A person was considered an online panel member from the moment that they filled out the first online survey that contains questions on the participants' personal profile, including socio-demographic characteristics and key substantive survey variables, such as voting behavior.

The GIP recruited new panel members in 2012 and 2014. The samples were drawn independently of each other and the sampling and recruitment procedures were almost identical. For the analyses in this paper, I therefore pool the two samples, except for the logistic regression model where I include a dummy variable for the recruitment year as a control variable.

I conduct all analyses separately for the face-to-face recruitment interview and the online profile survey to investigate whether auxiliary data predict survey response in either of the two recruitment steps.

4.2 The auxiliary data

The auxiliary data that I use in the analyses stem from various sources. Table 1 provides an overview of the variables, a brief description of the data, and information on the amount of missing data per variable.

Table 1: Overview of the auxiliary data, including brief data description and percentage of missing data

Variable	Description	Missing data (in %)
Sampling frame data	Official data; provided by the sampled municipalities	
Geographic region	East versus west Germany	0.00
Urbanity (admin.)	Less than 50,000 inhabitants, 50,000 to 500,000 inhabitants, more than 500,000 inhabitants	0.00
Urbanity (pol.)	Less than 20,000 inhabitants, 20,000 to 100,000 inhabitants, more than 100,000 inhabitants	0.00
Interviewer observations	Collected by the interviewers during fieldwork	
Intercom	Yes ; no	5.47
Type of building	Single housing unit versus apartment building	5.12
Building condition	Bad, moderate, good	5.23
Social class	Low, middle, high	9.23
Microm data	Commercial micro-geographic area data; from mixed and partly intransparent data sources	
Unemployment rate	Continuous variable	0.09
Exclusive livingenvironment	Yes ; no	3.22
Percentage academics	Continuous variable	3.22
Households in the street	Continuous variable	0.19
Percentage home ownership	Continuous variable	0.09
Percentage Roman Catholics	Continuous variable	0.01
Percentage Protestants	Continuous variable	0.01
INKAR data	Official micro-geographic area data; predominantly from the German Census and German Mikrozensus	
Percentage people aged 0-5	Continuous variable	0.00
Total balance of people that move to or from the district	Continuous variable	0.00
Percentage immigrants	Continuous variable	0.00
Aggregate household income	Continuous variable	0.00

4.2.1 Sampling frame data

The sampling frame data in my analyses are official data from the area database that is used to draw the GIP primary sampling units from. The sampling frame data include the geographic region (east

versus west), the degree of urbanity operationalized as administrative districts²² (less than 50,000 inhabitants, 50,000 to 500,000 inhabitants, more than 500,000 inhabitants), and the degree of urbanity operationalized as political governmental districts²³ (less than 20,000 inhabitants, 20,000 to 100,000 inhabitants, more than 100,000 inhabitants). The latter two measure population density in differently defined types of geographic areas. There is no missing data on any of the sampling frame variables.

4.2.2 Interviewer observations

The interviewers collected the interviewer observations during the fieldwork of the face-to-face recruitment interview. The interviewers were instructed to record the observations in the contact forms for all households in the gross sample units, so that the data are available for the respondents as well as the nonrespondents. The data include the presence of an intercom (yes, no), the type of building (single housing unit, apartment building), the building condition (bad, moderate, good), and the social class of the sampled household (low, middle, high). There is a substantial amount of missing data on all of the interviewer observations in the GIP, ranging between 5.12% on the type of building and 9.23% on social class.

4.2.3 Microm data

Microm data are micro-geographic area data provided by *Microm Consumer Marketing*, a commercial data vendor. The Microm data were linked to the sampled addresses of the GIP and are based on a variety of commercial and official sources such as Deutsche Telekom and the German Federal Employment Agency. In the documentation of the data, there is little detailed information on the data sources and adjustments²⁴. The variables I use are the unemployment rate (continuous), exclusive living environment (yes versus no), percentage of academics (continuous), number of households in the street (continuous), percentage of home ownership (continuous), percentage of

²² BIK; see <https://www.bik-gmbh.de/cms/basisdaten/bik-regionenfor> for more information

²³ GKPOL;

see http://www.bbsr.bund.de/BBSR/DE/Raumbeobachtung/Raumabgrenzungen/StadtGemeindetyp/StadtGemeindetyp_node.html for more information.

²⁴ See https://www.microm.de/fileadmin/media/document/Handbuch_Daten_2015_DE.pdf for more information.

Roman Catholics (continuous), and percentage of Protestants (continuous). Two of the variables are aggregated to up to four buildings (exclusive living environment, percentage of academics), two variables are aggregated to the street level (number of households, percentage of homeownership), two variables are aggregated to the municipality-level (percentage of Roman Catholics, percentage of Protestants), and one variable is aggregated to the postal-code-level (unemployment rate). There is a small to moderate amount of missing data on the Microm data, ranging between 0.01% on the percentage of Roman Catholics as well as the percentage of Protestants and 3.22% on exclusive living environment as well as the percentage of academics.

4.2.4 INKAR data

INKAR (Indicators and Maps for City and Spatial Research; Bundesinstitut für Bau-, Stadt- und Raumforschung, 2015) data are official micro-geographic area data aggregated to the municipality-level. They are predominantly based on the German Census and German Mikrozensus and can be downloaded free of charge from an Internet platform²⁵ that is run by the German Federal Institute for Research on Building, Urban Affairs, and Spatial Development (BBSR) of the German Federal Office for Building and Regional Planning (BBR). In the documentation of the data there is rich information on the data sources and adjustments. The variables I used are the percentage of people aged 0 to 5 (continuous), the total balance of people that move to or from the district (continuous), the percentage of immigrants (continuous), and the aggregate household income (continuous). There is no missing data on any of the INKAR data.

4.3 Methods

The analyses start with an examination of the data quality of the auxiliary data in terms of item missingness. For each auxiliary variable, I examine whether there are significant differences between GIP nonrespondents and respondents regarding the proportion of missing values. I use χ^2 -statistics to determine whether the Missing Completely at Random (MCAR; see Rubin & Little, 2002)

²⁵See www.inkar.de.

assumption holds. If the MCAR assumption does not hold the value of the auxiliary data for assessing and adjusting for misrepresentation is compromised.

My next analyses focus on assessing the utility of the different types of auxiliary data for predicting survey response. I first explore to what extent each auxiliary variable correlates with survey response using Pearson's correlation calculated from bivariate logistic regression models. Then, I examine the predictive power as well as the significance of coefficients for each type of auxiliary data in multivariate logistic regression models on response in the GIP recruitment interview and the GIP profile survey. I apply complete-case analyses in all of the regression models. Due to the missing data on some auxiliary variables, there are therefore 1,795 observations (12.92%) missing from the multivariate survey response models.

To differentiate between panel recruitment steps, I conduct the analyses separately on the face-to-face recruitment interview and the subsequent online profile survey.

5. Results

5.1 Missing data

Table 2 displays the proportion of missing values among nonrespondents and respondents on each of the variables that has at least some missing values separately for each GIP recruitment step. In addition, Table 2 provides χ^2 -statistics for the comparison between the proportion of missing values among nonrespondents and respondents.

Overall, I find moderately large proportions of missing values on the interviewer observations as well as the Microm data. For some of the interviewer observations, the proportion of missing values is significantly higher for nonrespondents than for respondents. In the recruitment interview, there is one variable where the proportion of missing values is significantly higher for nonrespondents than for respondents: the interviewer observation on social class (5.36% versus 3.87% with $\chi^2=38.68$, $df=1$, and $p=0.00$). In the profile survey, I find additional variables for which the proportion of missing

values differs between nonrespondents and respondents. There are significantly higher proportions of missing values among nonrespondents than among respondents on all of the interviewer observations: intercom (4.63% versus 0.84% with $\chi^2=16.39$, $df=1$, and $p=0.00$), type of building (4.33% versus 0.79% with $\chi^2=15.01$, $df=1$, and $p=0.00$), building condition (4.41% versus 0.82% with $\chi^2=14.14$, $df=1$, and $p=0.00$), and social class (7.99% versus 1.24% with $\chi^2=51.60$, $df=1$, and $p=0.00$).

Table 2: Proportion of missing values and χ^2 -statistic for the difference in proportion of missing values between nonrespondents (NR) and respondents (R) by auxiliary variable and GIP recruitment step (degrees of freedom (df) and p-values in parentheses; sampling frame data and INKAR data do not have any missing values)

	Missing data: Recruitment interview			Missing data: Profile survey		
	% NR	% R	χ^2	% NR	% R	χ^2
Interviewer observations						
Intercom	2.74	2.73	0.03 ($df=1$; $p=0.87$)	4.63	0.84	16.39***($df=1$; $p=0.00$)
Type of building	2.61	2.52	0.30 ($df=1$; $p=0.58$)	4.33	0.79	15.01***($df=1$; $p=0.00$)
Building condition	2.63	2.61	0.04 ($df=1$; $p=0.84$)	4.41	0.82	14.14*** ($df=1$; $p=0.00$)
Social class	5.36	3.87	38.68***($df=1$; $p=0.00$)	7.99	1.24	51.60***($df=1$; $p=0.00$)
Microm data						
Unemployment rate	0.06	0.04	0.71 ($df=1$; $p=0.40$)	0.06	0.03	0.71 ($df=1$; $p=0.40$)
Exclusive living environment	1.59	1.63	0.05 ($df=1$; $p=0.83$)	2.54	0.68	0.00 ($df=1$; $p=0.99$)
Percentage academics	1.59	1.63	0.05 ($df=1$; $p=0.83$)	2.54	0.68	0.00 ($df=1$; $p=0.99$)
Number hhlds. in the street	0.08	0.12	0.90 ($df=1$; $p=0.34$)	0.16	0.04	0.12 ($df=1$; $p=0.73$)
Percentage home ownership	0.06	0.04	0.71 ($df=1$; $p=0.40$)	0.06	0.03	0.71 ($df=1$; $p=0.40$)
Percentage Roman Catholic	0.01	0.00	2.01 ($df=1$; $p=0.16$)	0.01	0.00	0.54 ($df=1$; $p=0.46$)
Percentage Protestant	0.01	0.00	2.01 ($df=1$; $p=0.16$)	0.01	0.00	0.54 ($df=1$; $p=0.46$)

Note: *** $p<0.001$; ** $p<0.01$; * $p<0.05$

5.2 Correlations with survey response

Table 3 displays the correlation between each auxiliary variable and survey response at each GIP recruitment step. Generally, I find that correlations are small ranging from 0.00 to 0.18. By far the highest correlations with survey response can be found on the social class variable, although even

this correlation is rather weak. In the recruitment interview, no auxiliary variable correlates with survey response to any substantial degree (i.e. the correlation coefficients are smaller than 0.10). In the profile survey, only social class correlates with survey response, although only weakly ($r=0.18$).

Table 3: Correlation of auxiliary variables with survey response by auxiliary variable and GIP recruitment step (Pearson's correlation coefficients calculated from bivariate logistic regression models)

Variable	Pearson's r: Recruitment interview	Pearson's r: Profile survey
Sampling frame data		
Geographic region	0.00	0.01
Urbanity (admin.)	0.06	0.02
Urbanity (pol.)	0.06	0.02
Interviewer observations		
Intercom	0.05	0.01
Type of building	0.06	0.03
Building condition	0.05	0.08
Social class	0.08	0.18
Microm data		
Unemployment rate	0.02	0.05
Exclusive living environment	0.01	0.01
Percentage academics	0.01	0.01
Households in the street	0.04	0.04
Percentage home ownership	0.06	0.06
Percentage Roman Catholics	0.04	0.04
Percentage Protestants	0.00	0.00
INKAR data		
Percentage people aged 0-5	0.03	0.01
Total balance of people that move to or from the district	0.03	0.02
Percentage immigrants	0.01	0.02
Aggregate household income	0.04	0.01

5.3 Survey response models

Table 4 displays the results of the multivariate logistic regressions of the auxiliary variables on survey response separately for the two GIP recruitment stages. Overall, I find very low Pseudo R-squared for the survey response models even though many coefficients are significant.

Table 4: Logistic regression models of all auxiliary data (including dummy variable on the GIP recruitment phase) on the propensity to respond in the recruitment interview and the profile survey of the GIP (***) p<0.001; ** p<0.01; * p<0.05; standard errors in parentheses)

	Recruitment interview	Profile survey
Constant	-1.37*** (0.36)	-2.83*** (0.46)
Sampling Frame data		
Region: east (ref.: west)	0.70*** (0.11)	0.59*** (0.14)
Urbanity (admin.): 50k-500k inhabitants (ref.: <50k inhabitants)	-0.04 (0.05)	0.03 (0.06)
Urbanity (admin.): >500k inhabitants (ref.: <50k inhabitants)	-0.05 (0.08)	-0.01 (0.10)
Urbanity (pol.): 20k- 100kinhabitants (ref.: <20k inhabitants)	-0.14** (0.05)	-0.06 (0.06)
Urbanity (pol.): >100,000 inhabitants (ref.: <20k inhabitants)	-0.20** (0.07)	-0.00 (0.09)
Interviewer observations		
Intercom: yes (ref.: no)	-0.17*** (0.04)	-0.06 (0.05)
Type of building: Single housing unit (ref.: multiple units)	0.09 (0.05)	-0.01 (0.06)
Building condition: moderate (ref.: bad)	0.36*** (0.10)	0.38* (0.15)
Building condition: good (ref. bad)	0.46*** (0.11)	0.44** (0.16)
Social class: middle (ref.: low)	0.23*** (0.04)	0.84*** (0.05)
Social class: high (ref.: low)	0.74*** (0.08)	1.27*** (0.08)
Microm data		
Unemployment rate	0.02** (0.00)	0.00 (0.01)
Exclusive living environment: yes (ref. no)	0.03 (0.04)	0.01 (0.05)
Percentage academics	-0.00 (0.04)	0.05 (0.05)
Households in the street	0.00 (0.00)	0.00 (0.00)
Percentage home ownership	-0.00 (0.00)	0.00 (0.00)
Percentage Roman Catholics	0.01*** (0.00)	0.01*** (0.00)
Percentage Protestants	0.01*** (0.00)	0.01** (0.00)
INKAR data		
Percentage people aged 0-5	-0.04 (0.04)	-0.00 (0.05)

Table 4 (continued)

Total balance of people that move to or from the district	0.00 (0.00)	0.01 (0.00)
Aggregated household income	0.00 (0.00)	-0.00 (0.05)
Percentage immigrants	0.01 (0.01)	0.00 (0.01)
GIP recruitment year		
Recruitment year: 2012 (ref.: 2014)	0.34*** (0.04)	-0.33*** (0.05)
Pseudo R2	0.02	0.04
N	12098	12098

Note: *** p<0.001; ** p<0.01; * p<0.05

In the recruitment interview, the Pseudo R-Squared of the survey response model is 0.02. Regression coefficients are significant for two sampling frame variables (geographic region and urbanity (pol.)), most interviewer observations (intercom, building condition, and social class), and three Microm variables (unemployment rate, percentage Roman Catholics, and percentage Protestants).

In the profile survey, the Pseudo R-Squared of the survey response model is 0.04. Regression coefficients are significant for one sampling frame variable (geographic region), two interviewer observations (building condition and social class), and two Microm variables (percentage Roman Catholics and percentage Protestants).

When comparing the survey response models of the two GIP recruitment steps, I find that the Pseudo R-Squared of the survey response models are very low in the recruitment interview as well as the profile survey. In addition, there are three auxiliary variables (urbanity (pol.), Intercom, and unemployment rate) that have a significant effect in the survey response model in the recruitment interview but not in the profile survey. Furthermore, I find that the GIP recruitment year has a significant effect on survey response at both recruitment stages.

6. Discussion

This study assesses the utility of auxiliary data for predicting recruitment success in the GIP, a probability-based online panel in Germany. The auxiliary data I examined include sampling frame data, interviewer observations, and commercial as well as official micro-geographic area data. I assessed the utility of these types of auxiliary data for predicting survey response in the GIP face-to-face recruitment interview as well as the subsequent online profile survey.

First, I examined whether auxiliary data are systematically missing by survey response. I found that there are systematically more missing values for the nonrespondents than for the respondents on social class in the recruitment interview as well as all interviewer observations at the profile survey. The missingness mechanism on the interviewer observations can, therefore, not be considered MCAR. To correct for this bias is imperative to accurately estimate survey response using the interviewer observations.

To correct the missingness bias in the interviewer observations in estimations of survey response, it would, however, be necessary to have data available that correlate highly with the interviewer observations and survey response. In my study, the only data available for nonrespondents as well as respondents are the other auxiliary data. In addition, the aim of using the auxiliary variables is to predict survey response and then to be able to correct for potential nonresponse bias in the data. If there is already a missingness bias in the auxiliary variables, this might defeat the purpose of using the auxiliary data for representativeness assessments and nonresponse adjustments.

Second, I examined the correlations of each of the auxiliary variables with survey response. I found that the auxiliary data are at most weakly correlated with survey response. None of the auxiliary data correlates with survey response to any substantial degree in the recruitment interview and only social class correlates with survey response to some relevant degree in the profile survey.

Next, I assessed the predictive power of the auxiliary variables as well as the significance of their coefficients in logistic regression models on survey response. I found that the predictive power of the

auxiliary variables is very low in survey response models for both GIP recruitment steps. The predictive power of the multivariate logistic regression models of the auxiliary variables on survey response is low in the GIP recruitment interview (Pseudo R-Squared=0.02) as well as the profile survey (Pseudo R-Squared=0.04). However, in the recruitment interview the coefficients of most sampling frame data, most interviewer observations, and some Microm data are significant. In the profile survey, the coefficients of one sampling frame variable, some interviewer observations, and some Microm data are significant. None of the INKAR data coefficients is significant in the models. The results from the logistic regression models on survey response show that no type of auxiliary data contributes substantially to the prediction of recruitment success in the GIP at the two recruitment steps even though some coefficients are significant in the regression models.

To sum up, I find that all of the auxiliary data in my study have problems: the sampling frame data are limited to a few geographic characteristics, the interviewer observations contain relatively high proportions of missing values and this missingness is systematically related to survey response, the Microm data are intransparent and also contain some item missings, and the INKAR data are aggregated to a high level. None of the auxiliary variables is substantially correlated with survey response (with the exception of the interviewer observation social class for response to the online profile survey). Therefore, no single group of auxiliary data has substantial predictive power in the survey response models across the GIP recruitment steps.

I conclude from this study that no type of auxiliary data adds substantial value to predicting the recruitment success of the GIP. One potential reason for this might be that there is little nonresponse bias in the GIP recruitment. Other publications on the GIP recruitment, however, indicate some misrepresentation. When comparing survey estimates to benchmark data from official statistics, Blom et al. (2016) find, for example, that there is some misrepresentation on socio-demographic characteristics, such as age and education.

My study focused on the utility of auxiliary data for predicting survey response across the recruitment steps of a probability-based online panel. Because this study focuses on the level of the sampled households, correlations with substantive survey findings on the individual level are not considered here. Future research may extend this analysis and look into the association of auxiliary data with substantive survey variables. The auxiliary data I examined in my study might, for instance, still be useful for nonresponse adjustment weights if they are highly correlated with substantive survey variables.

Generally, auxiliary data can meaningfully contribute to survey operational tasks, such as eligibility screening and fieldwork monitoring, as well as post-survey nonresponse adjustments, such as weighting and imputations. In light of decreasing response rates and rising concerns about nonresponse bias and cost efficiency, these tasks are gaining in importance. However, I conclude from my study that the currently available auxiliary data should be used with caution because they might lack in general quality and they might not be predictive of survey response. These problems with the auxiliary data can compromise their use in survey operational tasks and adjustment procedures. My study shows that concerns about quality and predictive value of auxiliary data are justified and that the problems apply to a variety of commonly used types of auxiliary data. I therefore conclude from this research that the search for high quality auxiliary data that is predictive of survey response needs to continue to be able to satisfy the need for such auxiliary data in survey practice.

Literature

Biemer, P. P., Peytchev, A. (2013). Using Geocoded Census Data for Nonresponse Bias Correction: An Assessment. *Journal of Survey Statistics and Methodology*, 1(1), 24–44.

Blom, A. G., Gathmann, C., & Krieger, U. (2015). Setting up an online panel representative of the general population: The German Internet Panel. *Field methods*, 27(4), 391-408.

Blom, A. G., Herzing, J. M., Cornesse, C., Sakshaug, J. W., Krieger, U., & Bossert, D. (2016). Does the recruitment of offline households increase the sample representativeness of probability-based online panels? Evidence from the German internet panel. *Social Science Computer Review*, 0894439316651584.

Bundesinstitut für Bau-, Stadt- und Raumforschung (2015) *INKAR: Indikatoren und Karten zur Raum- und Stadtentwicklung*. © BBSR Bonn 2015. Bonn: Bundesamt für Bauwesen und Raumordnung. (Available from <http://www.inkar.de>.)

European Social Survey. (2012). ESS6 Source Contact Forms. Retrieved from http://www.europeansocialsurvey.org/docs/round6/fieldwork/source/ESS6_source_contact_forms.pdf.

Groves, R. M., & Heeringa, S. G. (2006). Responsive design for household surveys: tools for actively controlling survey errors and costs. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 169(3), 439-457.

Hall, J. (2008). Sampling Frame. In: Paul Lavrakas (ed.) *Encyclopedia of Survey Research Methods*. 790-791.

HRS. (2014). HRS 2014 Final Release Codebook. Retrieved from: http://hrsonline.isr.umich.edu/modules/meta/2014/core/codebook/h14_00.html.

Kirgis, N., & Lepkowski, J. (2010). A management model for continuous data collection: Reflections from the National Survey of Family Growth, 2006–2010. *NSFG Paper*, (10-011). Retrieved from <https://www.psc.isr.umich.edu/pubs/pdf/ng10-011.pdf>.

Kreuter, F., Olson, K., Wagner, J., Yan, T., Ezzati-Rice, T. M., Casas-Cordero, C., Lemay, M., Peytchev A., Groves, R.M., & Raghunathan, T. E. (2010). Using proxy measures and other correlates of survey outcomes to adjust for non-response: examples from multiple surveys. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 173(2), 389-407.

Matsuo, H., Billiet, J., & Loosveldt, G. (2010). Response-based quality assessment of ESS Round 4: Results for 24 countries based on contact files. Retrieved from: https://www.europeansocialsurvey.org/docs/round4/methods/ESS4_response_based_quality_assessment_e02.pdf.

Olson, K. (2013). Paradata for nonresponse adjustment. *The ANNALS of the American Academy of Political and Social Science*, 645(1), 142-170.

Pasek, J., Jang, S. M., Cobb III, C. L., Dennis, J. M., & Disogra, C. (2014). Can marketing data aid survey research? Examining accuracy and completeness in Consumer-File data. *Public Opinion Quarterly*, 78(4), 889-916.

Rubin, D. B., & Little, R. J. (2002). *Statistical analysis with missing data*. John Wiley & Sons.

Sampson, R. J., & Raudenbush, S. W. (1999). Systematic social observation of public spaces: A new look at disorder in urban neighborhoods. *American journal of sociology*, 105(3), 603-651.

Särndal, C. E. (2011). The 2010 Morris Hansen Lecture. Dealing with survey nonresponse in data collection, in estimation. *Journal of Official Statistics*, 27(1), 1.

Sakshaug, J., Antoni, M., & Sauckel, R. (2017). The Quality and Selectivity of Linking Federal Administrative Records to Respondents and Nonrespondents in a General Population Sample Survey of Germany. *Survey Research Methods* 11(1), 63-80.

Schouten, B., Cobben, F., & Bethlehem, J. (2009). Indicators for the representativeness of survey response. *Survey Methodology*, 35(1), 101-113.

Schouten, B., Calinescu, M., & Luiten, A. (2013). Optimizing quality of response through adaptive survey designs. *Survey Methodology*, 39(1), 29-58.

Schouten, B., Shlomo, N. and Skinner, C. (2011) Indicators for Monitoring and Improving Representativeness of Response. *Journal of Official Statistics*, 27(2), 1-24.

Schouten, B., Bethlehem, J., Beullens, K., Kleven, Ø., Loosveldt, G., Luiten, A., Rutar, K., Shlomo, N., & Skinner, C. (2012). Evaluating, Comparing, Monitoring, and Improving Representativeness of Survey Response Through R-Indicators and Partial R-Indicators. *International Statistical Review*, 80(3), 382-399.

Schouten, B., Cobben, F., Lundquist, P. and Wagner, J. (2016), Does more balanced survey response imply less non-response bias? *Journal of the Royal Statistical Society, A*, 179, 727–748.

Sinibaldi, J., Trappmann, M., & Kreuter, F. (2014). Which is the better investment for nonresponse adjustment: Purchasing commercial auxiliary data or collecting interviewer observations? *Public Opinion Quarterly*, 78(2), 440-473.

Wagner, J. R. (2008). *Adaptive survey design to reduce nonresponse bias* (Doctoral dissertation, University of Michigan). Retrieved from <https://search.proquest.com/docview/304573555?pq-origsite=gscholar>.

Wagner, J. (2010). The fraction of missing information as a tool for monitoring the quality of survey data. *Public Opinion Quarterly*, 74(2), 223-243.

Wagner, J., West, B. T., Kirgis, N., Lepkowski, J. M., Axinn, W. G., & Ndiaye, S. K. (2012). Use of paradata in a responsive design framework to manage a field data collection. *Journal of Official Statistics*, 28(4), 477.

West, B. (2010). An Examination of the Quality and Utility of Interviewer Estimates of Household Characteristics in the National Survey of Family Growth. NSFG Working Paper, (10-009). Retrieved from <https://www.psc.isr.umich.edu/pubs/pdf/ng10-009.pdf>.

West, B. T., Kreuter, F., & Trappmann, M. (2014). Is the collection of interviewer observations worthwhile in an Economic Panel Survey? New evidence from the German Labor Market and Social Security (PASS) study. *Journal of Survey Statistics and Methodology*, 2(2), 159-181.

West, B. T., & Kreuter, F. (2013). Factors affecting the accuracy of interviewer observations: Evidence from the National Survey of Family Growth. *Public Opinion Quarterly*, 77(2), 522-548.

West, B. T., & Little, R. J. (2013). Non-response adjustment of survey estimates based on auxiliary variables subject to error. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 62(2), 213-231.

West, B. T., Wagner, J., Hubbard, F., & Gu, H. (2015). The utility of alternative commercial data sources for survey operations and estimation: Evidence from the National Survey of Family Growth. *Journal of Survey Statistics and Methodology*, 3(2), 240-264.

Paper 4

Response quality in nonprobability and probability-based online panels

Paper 4

Response quality in nonprobability and probability-based online panels²⁶

Abstract

Recent years have seen a growing number of studies investigating the accuracy of nonprobability online panels; however, response quality in nonprobability online panels has not yet received much attention. To fill this gap, we investigate response quality in a comprehensive study of seven nonprobability online panels and three probability-based online panels, which were all collected in Germany during the same fieldwork period. Three response quality indicators typically associated with survey satisficing are assessed: straight-lining in grid questions, item nonresponse, and midpoint selection in visual design experiments. Our results show that there is significantly more straight-lining in the nonprobability online panels. However, contrary to our expectations, there is no generalizable difference between nonprobability online panels and probability-based online panels with respect to item nonresponse. Finally, neither respondents in nonprobability online panels, nor respondents in probability-based online panels are significantly affected by the visual design of the midpoint of the answer scale.

Keywords

Response quality, nonprobability sample, probability-based sample, online panel, satisficing

Acknowledgements

The authors gratefully acknowledge support from the Collaborative Research Center (SFB) 884 “Political Economy of Reforms” (projects A8 and Z1), funded by the German Research Foundation (DFG) and from the GESIS Panel, funded by the German Federal Ministry of Education and Research.

²⁶ This paper is joint work with Annelies Blom.

In addition, we would like to thank Daniela Ackermann-Piek and Susanne Helmschrott for help with preparing the data as well as Christian Bruch and Barbara Felderer for valuable feedback on early analyses.

1. Introduction

The past decade has seen increasing debate about the quality of nonprobability online panels. This debate has primarily circled around whether or not these panels provide representative sets of respondents (see for example Smith 2003; Malhotra and Krosnick 2007; Loosveldt and Sonck 2008; Chang and Krosnick 2009; Baker et al. 2010; Yeager et al. 2011; Legleye et al. 2015; Pasek 2016). The apprehension of some researchers is that a biased subgroup of the population self-selects into nonprobability online panels (Malhotra and Krosnick 2007; Loosveldt and Sonck 2008; Chang and Krosnick 2009; Yeager et al. 2011; Pasek 2016; and Pennay et al. 2016). Others argue that nonprobability online panels can accurately reflect the target population, especially after weighting (see Wang et al. 2015; Goel, Obeng, and Rothschild 2015; and Kennedy, Keeter, and Weisel 2016). However, while the number of publications on nonprobability panel accuracy is increasing, less attention has been paid to response quality in nonprobability panels (for notable exceptions see Chang and Krosnick 2009; Greszki, Meyer, and Schoen 2014; and Hillygus, Jackson, and Young 2014).

This is surprising, because one might argue that, since their respondents participate in the panel mainly for monetary reasons (see for example GreenBook 2017), nonprobability online panel respondents may be less committed to the panel in terms of response quality. The focus on monetary rewards among the respondents is encouraged by the nonprobability online panel advertising industry that recruits the panel members. With advertising slogans like “Earn Cash With Quick Paid Surveys!” (www.quickpaysurvey.com), “Make Money Online With Paid Surveys” (www.cashcrate.com) or “Take surveys for cash” (www.takesurveysforcash.com) advertisers try to attract as many internet users as possible. Being attracted to the online panel by the promise of easy money for little effort, nonprobability online panel respondents may show different care in answering survey questions than respondents that were recruited into an online panel by probability-based offline recruitment methods, independent of their socio-demographic characteristics. And consequently, response quality might differ between nonprobability and probability-based online panel respondents.

In this paper, we investigate whether there is a difference in response quality between survey participants in nonprobability online panels and probability-based online panels. For this purpose we look into three indicators that are often associated with response quality in the context of survey satisficing: straight-lining, item nonresponse, and midpoint selection.

2. Respondent motivation and survey satisficing

Most social research based on survey data relies on the assumption that respondents answer the survey questions to the best of their ability. This requires the respondents to carefully carry out all the cognitive steps involved in answering a survey question. According to Tourangeau, Rips, and Rasinski (2000) cognitive response processing consists of four steps: question comprehension, information retrieval, judgment and estimation, and reporting an answer. Data analysts typically assume that all four response steps were carefully carried out by all respondents, i.e. that respondents optimize survey responses. Respondents, however, sometimes take short-cuts through the optimal cognitive response process. This behavior is called satisficing (Krosnick and Alwin 1987).

Krosnick (1991) defines two types of satisficing: weak satisficing and strong satisficing. Weak satisficing occurs when respondents execute all cognitive steps that are necessary to arrive at a response, but they do so only superficially. They might, for instance, read a question text carefully but skip the accompanying instruction text. Alternatively, when presented with a list of answer options they might choose the first option that approximately fits their opinion without considering further answer options that might fit their opinion even better. Because weak satisficers carry out the response steps superficially, unmeaningful decision-making cues, like the visual design of an answer scale, can influence their responses.

While weak satisficers still carry out all four cognitive response steps, even if they do so superficially, strong satisficers do not or only partially carry out the response steps. Respondents might, for instance, process only just enough information to arrive at a response that they consider generally reasonable without reading and considering the question carefully, or without searching their

memories and retrieving the relevant information. This strategy results in either no responses at all, unmeaningful answers, nonsubstantive answers, or undifferentiated answers.

According to satisficing theory, there is a continuum of cognitive thoroughness of responses with perfectly optimized responses at one end of the continuum and strongly satisfied responses at the other end (see Krosnick 1991). Respondents can be in different positions on the continuum, with some being generally thorough in all of their responses and others being generally less careful in their responses (for empirical proof see Krosnick 1992; Narayan and Krosnick 1996; Knäuper 1999; Malhotra 2008; and Kaminska, McCutcheon, and Billiet 2010). In addition to this interpersonal variation, there can be intra-personal variation in the level of satisficing observed during an interview, for example when respondents fatigue during a long interview.

According to Krosnick (1991), there are three factors that largely influence the occurrence and degree of satisficing: task difficulty, respondent ability, and respondent motivation. Task difficulty refers to the cognitive effort needed to answer a question optimally. The task difficulty depends on the complexity of a question and of the information asked for. When a question contains many words, long words, and/or uncommon or ambiguous terms, satisficing is more likely to occur than when questions are short, precise, and easy to comprehend (for empirical proof see for example Alwin and Krosnick 1991). Similarly, when a question asks respondents to evaluate multiple items or answer on a long, unlabeled scale, satisficing is more likely than when a question asks for the evaluation of only one item and provides few fully labeled answer options (see also Krosnick and Berent 1993; and Krosnick 1999).

Respondent ability refers to the competences and skills involved in answering survey questions. It includes cognitive abilities, such as cognitive sophistication, the amount of practice in thinking about a topic, and attitude strength, as well as practical abilities, such as the ability to process and communicate answers. In self-administered questionnaires, respondent ability also includes reading and writing skills. Furthermore, in web surveys, computer literacy and technological skills necessary

to start and navigate through the survey are important aspects of respondent ability (for empirical proof on satisficing in web surveys see for example Toepoel et al. 2009). Satisficing is more likely to occur when a respondent has low cognitive abilities and may therefore have problems comprehending the question (see also Krosnick and Alwin 1987; Krosnick 1992; Narayan and Krosnick 1996; and Kaminska et al. 2010).

The respondents' motivation determines how much effort they are willing to invest in answering a question. To some extent, respondent motivation is a personality characteristic that is related to a person's need for cognition (for information on the concept of need for cognition see Cacioppo and Petty 1982). Satisficing is more likely to occur if respondents have low need for cognition (see Kaminska et al. 2010). Respondents with high need for cognition have an intrinsic motivation to fill out the questionnaire carefully, because they generally enjoy thinking about questions and have fun expressing their opinions. High interest in a survey topic and the perceived importance of a survey can have a substantial impact on respondent motivation at the start of the survey (see also Groves, Presser, and Dipko 2004; Stoop 2005). However, respondent motivation may decrease over the course of the interview. Therefore, satisficing is more likely towards the end of a survey, especially in long surveys, than towards the start and in short surveys (see also Krosnick et al. 2002).

3. Measuring satisficing

Thus, according to Krosnick (1991), the combination of task difficulty, respondent ability, and respondent motivation explains the occurrence and amount of satisficing. In our analyses, we keep task difficulty constant across panels by including exactly the same questions across all of the panels. In addition, we aim to keep respondent ability constant across panels by applying the same weighting procedure to all samples and thus controlling for sample composition differences in socio-demographic characteristics.

Regarding respondent motivation, we expect the nonprobability panel participants to be motivated mainly by the monetary incentives that they receive in their respective panels. Thus, we expect them

to minimize effort in order to maximize their incentive-by-effort ratio. This expectation is supported by studies that demonstrate the importance of monetary incentives for nonprobability online panel members. For instance, in a study on the motives for joining nonprobability online panels, Keusch, Batinic, and Mayerhofer (2014:179) find that when asked to select all of their motives for participation, 40% of panelists indicated they “wanted to earn some extra money”. In addition, “monetary motives had the strongest correlation with survey participation” for nonprobability online panel participants (Keusch et al. 2014:185). Similarly, Sparrow (2006:5) finds that 52% of the new members of a nonprobability online panel participate because it is “an enjoyable way to earn money” as opposed to 20% who join because they “thought they would be interested in the topics covered”, and 19% who “enjoy answering questions”.

The probability-based online panelists are less driven by monetary incentives. For instance, analyzing survey data on the respondents’ most important reason to participate in the Dutch LISS Panel²⁷, a probability-based online panel of the general population in the Netherlands, we find that 16.4% of the respondents participate in the panel because they “find it important to contribute to science”. The financial reward is only stated as the most important reason for participating by 15.2% of respondents. Furthermore, for the German Internet Panel (GIP²⁸), a probability-based online panel of the German general population, we observe that about 13% of panelists even waive their incentive and choose to donate it to charity instead.

We, therefore, expect the share of people who care about giving optimized answers to be lower, and thus the amount of satisficing behavior to be higher, in the nonprobability online panels than in the probability-based online panels. To investigate differences in response quality across nonprobability and probability-based online panels, we examine three indicators of satisficing: straight-lining in grid questions, item nonresponse, and midpoint selection in a visual design experiment. In the following, we discuss the theoretical and empirical background in the literature of each of these indicators.

²⁷ www.lissdata.nl

²⁸ http://reforms.uni-mannheim.de/internet_panel

3.1 Straight-lining

The term straight-lining refers to the tendency of respondents to choose the same or a very similar answer option for each item in a grid (see Schonlau and Toepoel 2015). This phenomenon is sometimes also referred to as “non-differentiation” (see Malhotra, Miller, and Wedeking 2014). Straight-lining is a strong form of satisficing. It occurs in self-administered questionnaires, because the grid format provides a visual cue that triggers a specific type of cognitive shortcut: While for the first item of the grid respondents might still carry out all the cognitive steps necessary to arrive at an optimized response, the grid format suggests that the same answer will also be acceptable for the following items. Therefore, some respondents might abandon the full cognitive response processing in favor of a shortcut and give the same (or a very similar) answer to all other grid items. Research suggests that avoiding grid questions and asking each question separately instead, preferably with only one question per screen, can prevent straight-lining (see Couper, Traugott, and Lamias 2001; and Couper 2008).

3.2 Item nonresponse

Like straight-lining, item nonresponse is a form of strong satisficing. Respondents who choose this satisficing strategy skip one or all of the cognitive response steps. In a web survey, respondents might not read the question text carefully and click on the “next”-button instead, in order to get to the end of the questionnaire more quickly. If they read the question carefully, they might still not be willing or able to go through the necessary information retrieval or the judgement and estimation processes. There are two types of item nonresponse: question skipping (QS) and giving a nonsubstantive answer, i.e. answering “don’t know” (DK) or “don’t want to say” (DWS).

Nonsubstantive responses require some cognitive effort because respondents have to at least browse through the answer options or look for other visual design cues that lead them to the DK or DWS answer category. Relative to nonsubstantive responses, QS is a much stronger type of satisficing, because when choosing to skip a question altogether respondents do not engage in any kind of cognitive response process.

A potential drawback to using nonsubstantive answers as a satisficing indicator is that respondents may carry out all the necessary response steps and in the end still decide to choose a nonsubstantive response, for example because they honestly do not know the answer to a question (Converse 1974; Schuman and Presser 1981; and Sturgis, Roberts, and Smith 2014). With respect to “no opinion” answers, Krosnick et al. (2002), however, show that respondents with low levels of education are more likely to choose this answer option suggesting that people are more likely to choose “no opinion” responses when they perceive the processes of producing an optimal response as cumbersome. Furthermore, the authors find that the amount of “no opinion” answers increases with interview duration, suggesting that respondent motivation decreases towards the end of the interview resulting in less willingness to engage in the cognitive effort necessary to produce an optimal response. These findings are supported by further research (see for example Bradburn and Sudman 1988; Feick 1989; Gilljam and Granberg 1993; Fowler and Cannell 1996; and Holbrook, Green, and Krosnick 2003).

3.3 Midpoint selection

The midpoint of a scale provides a superficial visual cue in self-administered questionnaires. Following Krosnick’s (1991) reasoning, some respondents might satisfice by selecting the middle category while optimizing respondents are not influenced by this visual cue and instead choose the answer option that best represents their “true” answer after carrying out all the necessary cognitive steps. Specifically, Krosnick and Fabrigar (1997:147) argue that “many people [...] might select an offered midpoint because it provides an easy choice that requires little effort and is easy to justify”. On the continuum of cognitive thoroughness, midpoint selection can be interpreted as a weaker form of satisficing than straight-lining and item nonresponse when we assume that respondents go through the necessary response steps but are influenced in their decision by the visual design cue that the scale midpoint provides. Tourangeau, Couper, and Conrad (2004) show that, generally, the visual midpoint rather than the conceptual midpoint shapes the response distribution. Their explanation is that “the visual midpoint is seen as providing a benchmark, representing either the

conceptual midpoint of the scale or the most typical response” (Tourangeau et al. 2004:390). Following satisficing theory, it is likely that satisficing respondents choose the visual midpoint to reduce the cognitive effort needed to evaluate the other response options against this midpoint. This behavior would constitute a shortcut to the information retrieval and/or the judgement and estimation processes. Therefore, the selection of the visual midpoint may be used as an indicator of satisficing behavior (see also Kaminska et al. 2010; and Malhotra et al. 2014).

4. Online panel data

In this paper, we assess response quality among respondents in ten online panels (see Table 1). Three of them are probability-based online panels: the German Internet Panel (GIP), the GESIS Panel (GP), and one commercial probability-based online panel. The remaining seven are commercial nonprobability online panels. In each of the online panels, we fielded the same short multi-topic survey. In the following, we describe the various online panels and the study that we implemented in more detail.

Table 1: Panel characteristics

Panel ²⁹	Type	Sampling	Recruitment	Offline households	Fieldwork period	N
GIP	Academic	Probability	Face-to-face	Yes	1-31 March 2015	3075
GP	Academic	Probability	Face-to-face	Yes	18 Feb14-April 2015	2533
1	Commercial	Probability	Telephone	No	1-31 March 2015	1012
2	Commercial	Nonprobability	Online	No	5-18 March 2015	1038
3	Commercial	Nonprobability	Online	No	2-11 March 2015	999
4	Commercial	Nonprobability	Online	No	1-18 March 2015	1002
5	Commercial	Nonprobability	Online	No	2-16 March 2015	1000
6	Commercial	Nonprobability	Online	No	25 March-1 April 2015	1000
7	Commercial	Nonprobability	Online	No	3-9 March 2015	994
8	Commercial	Nonprobability	Online	No	5-11 March 2015	1000

4.1 The GIP

The GIP is based on a three-stage stratified probability area sample with subsequent face-to-face recruitment interviews. At the first sampling stage, a random sample of areas is drawn from a

²⁹Data collection in the GIP, GP, and Panel 1 was free of charge. The other panels are sorted by cost with Panel 8 being the most expensive panel.

database that covers all areas in Germany. Within each primary sampling unit (PSU), listers record every household along a predefined random route until they have listed 200 households. Subsequently, a random sample of households is drawn to be interviewed in face-to-face recruitment interviews. All age-eligible members of sampled households are invited to become online panelists (Blom, Gathmann, and Krieger 2015). Furthermore, the GIP covers individuals without computer and/or internet access by equipping them with the necessary devices (see Blom et al. 2017). All panel members are invited bi-monthly to participate in an online interview of about 20-25 minutes on a diversity of social, economic, and political topics. For the analyses in this paper, we used wave 16 of the GIP. It had a completion rate³⁰ of 69.8 percent and a cumulative response rate³¹ of 14.3 percent. Although the GIP covers the German population aged 16 to 75, we only use data from individuals that were aged 18 to 70 to make the data comparable across all panels in this study.

4.2 The GP

The GP is based on a two-stage stratified probability sample from population registers and subsequent face-to-face recruitment interviews. At the first sampling stage, the GP draws a random sample of areas from a database of municipalities in Germany. Then, the GP samples individuals from the local population registers within each of the sampled PSUs. Because 10 sampled municipalities refused to cooperate with the GP sampling request, these PSUs had to be substituted (see Bosnjak et al. 2017). Subsequently, the sampled individuals were contacted for face-to-face recruitment interviews. All interviewed individuals were invited to become panelists. Furthermore, the GP includes the offline population via paper-and-pencil mail surveys. Internet users that prefer to participate offline instead of online are also provided with paper-and-pencil mail surveys. All panel members of the GP are invited to participate in bi-monthly interviews of about 20 minutes on a wide

³⁰ Based on AAPOR Standard Definitions (AAPOR 2016); completion rate (COMR) = number of wave participants divided by the number of recruited panel members.

³¹ Based on AAPOR Standard Definitions (AAPOR 2016); cumulative response rate (CUMRR) = number of wave participants divided by the number of eligible persons

variety of topics. For the analyses in this paper, we used wave *ca* of the GP. It had a completion rate³² of 50.5 percent and a cumulative response rate³³ of 19.5 percent. The age range covered in the GP is 18 to 70 years.

In our study, we exclude the GP mail respondents because the potential mode effect might bias our results (see for example Green, Krosnick, and Holbrook 2001; and Holbrook et al. 2003 for evidence on differences in satisficing by mode).

4.3 Panel 1

Panel 1 is a commercial probability-based online panel. To recruit panel members, the panel draws its sample from Random Digit Dialing (RDD) telephone surveys conducted in-house by the same company. Individuals interviewed in an RDD telephone interview are subsequently invited to join the panel for regular online interviews. Panel 1 does not cover the offline population. For our study, Panel 1 drew a quota sub-sample from its probability-based respondent pool.

4.4 The nonprobability online panels: Panel 2 - Panel 8

For the recruitment of the nonprobability online panels, we published a call for tender in November 2014. The call for tender explained that we sought to implement a 10-minute questionnaire about traffic, politics, and health among 1,000 respondents that should be representative of the German population aged 18 to 70 years of age. The call further announced that the data were to be collected in March 2015. Regarding further design decisions (application of quotas, provision of weights, etc.) the panel providers were free to choose whichever approach they thought would provide the most representative data.

In response to our call, we received 17 tenders for conducting the specified survey. Out of these, 16 survey providers explicitly offered a sample representative of the general population in Germany

³² Completion rate = number of wave participants divided by the number of recruited panel members.

³³ Cumulative response rate = number of wave participants divided by the number of gross sample members.

aged 18 to 70. Seven providers were considered fit for our purpose as well as within reasonable budgetary limits. We therefore commissioned them with collecting our data. The costs quoted for conducting the wave analyzed in this paper plus two additional waves conducted in half-year intervals varied widely across the panels (see Appendix A for details). In the following, the nonprobability online panels are numbered from 2 to 8 in sequence of ascending costs.

5. Methods

To assess differences in the amount and type of satisficing between the nonprobability online panels and the probability-based online panels, we consider three satisficing indicators: straight-lining in grids, item nonresponse, and midpoint selection in a visual design experiment. Their operationalization is described in the following. Information on the question texts and answer scales of the questions we used in our analyses can be found in Appendix B.

5.1 Straight-lining

To assess straight-lining in our questionnaire, we implemented two psychological short scales with four items each in grid format. We define respondents as straight-liners if they choose the same answer category for every item on at least one of the two grid questions. Based on the literature described above, we test the following hypothesis:

H1: In nonprobability online panels a higher proportion of respondents chooses to straight-line on grid questions than in probability-based online panels.

To test this hypothesis, we compare the proportions of straight-liners and the respective confidence intervals around these estimates across the online panels. Furthermore, we apply χ^2 -tests to examine averages across the nonprobability and probability-based online panels.

5.2 Item nonresponse

Our questionnaire contained several possibilities for generating item nonresponse, all of which were implemented in the same way across all panels. In our analyses, we can differentiate between

different types of item nonresponse. At several questions, respondents were able to give nonsubstantive answers. At five questions we provided the nonsubstantive answer option “don’t know” (DK), at two questions we provided a “don’t want to say” (DWS) option, and at one question we provided both DK and DWS answer options. In addition, at each question, respondents were able to skip the question by clicking on the “next”-button, generating question skips (QS). Based on the literature described above, we test the following hypotheses:

H2: In nonprobability online panels a higher proportion of respondents chooses to not provide any (substantive) response to a question than in probability-based online panels.

H2a: In nonprobability online panels a higher proportion of respondents chooses to answer DK than in probability-based online panels.

H2b: In nonprobability online panels a higher proportion of respondents chooses to answer DWS than in probability-based online panels.

H2c: In nonprobability online panels a higher proportion of respondents chooses to skip a question than in probability-based online panels.

For each type of item nonresponse, we generated a dichotomous variable that operationalizes whether a respondent chose this type of item nonresponse at least once during the survey. In addition, we generated a variable INR that operationalizes whether a survey respondent ever chose any type of item nonresponse during the survey as an overview statistic of item nonresponse.

To test our hypotheses, we compare the proportions of each type of item nonresponse and the respective confidence intervals around these estimates across the online panels. Furthermore, we apply χ^2 -tests to examine averages across the nonprobability and probability-based online panels.

5.3 Midpoint selection

We implemented a visual design experiment in our questionnaire to investigate whether respondents answer consistently across different answer scales. Four experimental conditions were randomly assigned to respondents.

Condition 1: The conceptual midpoint was located at the visual midpoint.

Condition 2: The answer scale contained a conceptual midpoint but no visual midpoint.

Condition 3: The answer scale contained a visual midpoint but no conceptual midpoint.

Condition 4: The answer scale contained neither a conceptual nor a visual midpoint.

We conducted this experiment on two questions of our questionnaire and randomly assigned respondents independently at each question. The first question covered respondents' perceived health. The second question concerns respondents' opinion on environmental zones in cities³⁴.

We varied the presence of a *conceptual* midpoint by including a "neither nor" or "average" answer option in the scale versus excluding this conceptual midpoint. We vary the presence of a *visual* midpoint by leaving a gap between the substantive answer options and the "I don't know" option versus not leaving a gap. The respective answer scales are depicted in Table 2a and Table 2b.

³⁴ Environmental zones are areas in German cities, from which cars that emit high levels of respirable dust are banned.

Table 2a: Experimental conditions in the midpoint experiment on the 5-point scale

Conceptual & visual midpoint	Conceptual midpoint only	Visual midpoint only	No midpoint
Very good	Very good	Very good	Very good
Good	Good	Good	Good
Average	Average	Bad	Bad
Bad	Bad	Very bad	Very bad
Very bad	Very bad	DK	DK
DK	DK	DK	DK

Note:  conceptual midpoint;  visual midpoint.

Table 2b: Experimental conditions in the midpoint experiment on the 7-point scale

Conceptual & visual midpoint	Conceptual midpoint only	Visual midpoint only	No midpoint
Very good	Very good	Very good	Very good
Good	Good	Good	Good
Rather good	Rather good	Rather good	Rather good
Neither good nor bad	Neither good nor bad	Rather bad	Rather bad
Rather bad	Rather bad	Bad	Bad
Bad	Bad	Very bad	Very bad
Very bad	Very bad	DK	DK
DK	DK	DK	DK

Note:  conceptual midpoint;  visual midpoint.

As described in the literature above, we detect satisficing behavior when respondents are drawn to visual midpoints in self-completion questionnaires. Our experiment allows us to test the following hypotheses:

H3: In nonprobability online panels a higher proportion of respondents chooses an answer option located at the visual midpoint than in probability-based online panels.

H3a: The difference between the proportion of respondents choosing the conceptual midpoint when it matches the visual midpoint (condition 1) and the proportion of respondents choosing the conceptual midpoint when there is no visual midpoint (condition 2) will be higher in nonprobability online panels than in probability-based online panels.

H3b: The difference between the proportion of respondents choosing an answer option when it is located at the visual midpoint (condition 3) and the proportion of respondents choosing the same answer option when it is not located at the visual midpoint (condition 4) will be higher in nonprobability online panels than in probability-based online panels.

We apply χ^2 -tests to examine averages across the nonprobability and probability-based online panels. For each of the estimates in our analyses, we obtained bootstrapped standard errors by pooling results across 100 variance-covariance matrices.

6. Results

We examine whether there are significant differences in response quality between nonprobability online panels and probability-based online panels based on our hypotheses on the satisficing indicators described (straight-lining, item nonresponse, and midpoint selection in a visual design experiment).

6.1 Straight-lining

With regard to our hypothesis that a higher proportion of respondents chooses to straight-line on grid questions in nonprobability online panels than in probability-based online panels (H1), we

indeed find significantly more straight-lining in the nonprobability panels than in the probability-based panels (on average 13.4% and 6.5%, respectively; $\chi^2(1)=152.7, p=0.00$).

When investigating the online panels in detail (see Figure 1), we find that in each of the nonprobability online panels, straight-lining is considerably more prevalent (between 10.2% in Panel 2 and 16.2% in Panel 5) than in the probability-based online panels (between 3.1% in GP and 9.2% in GIP). In fact, two of the probability-based online panels (GP and Panel 1) show significantly less straight-lining than the best nonprobability online panel (Panel 2).

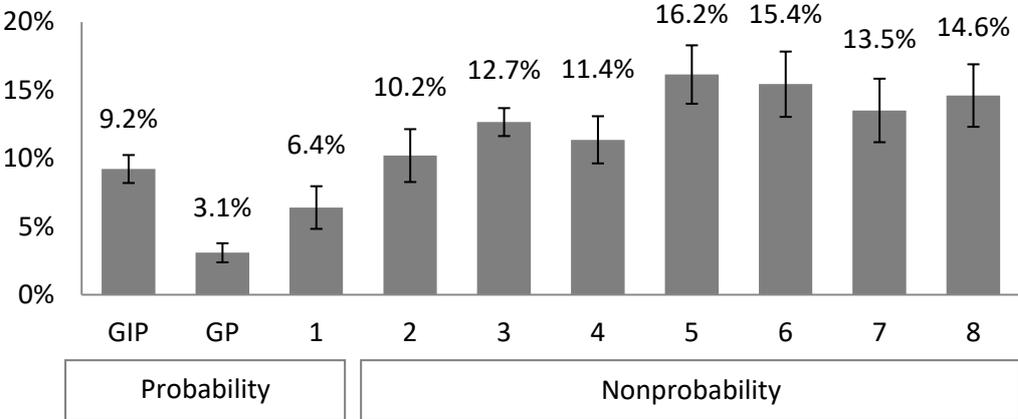


Figure 1: Proportion of straight-liners across panels (bars); bootstrapped 95% confidence intervals (spikes)

6.2 Item nonresponse

Regarding our hypothesis that a higher proportion of respondents chooses to not provide any (substantive) response to a question in nonprobability online panels than in probability-based online panels (H2), we find no generalizable evidence in support of our item nonresponse hypotheses across the three types of item nonresponse (DK, DWS, QS).

When examining INR as an overall measure of item nonresponse that operationalizes the broader hypothesis H2 (see INR bars in Figure 2), we find that, contrary to our expectations, a significantly lower proportion of respondents chose any type of item nonresponse at least once in the nonprobability online panels compared to the probability-based online panels (29.2% versus 31.7% respectively; $\chi^2(1)=9.7, p=0.00$). In the following, we explore the different types of item nonresponse (DK, DWS, QS) in more detail based on our hypotheses 2a to 2c.

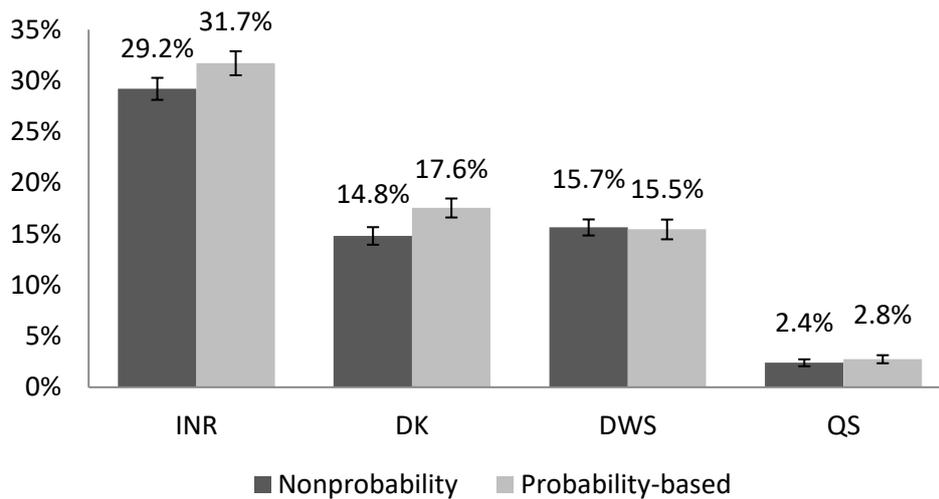


Figure 2: Proportion of item nonresponse in probability-based and nonprobability panels (bars); bootstrapped 95% confidence intervals (spikes)

With regard to our hypothesis that a higher proportion of respondents chooses to answer DK in nonprobability online panels than in probability-based online panels (H2a), we find that, contrary to our expectations, the nonprobability online panel respondents chose to answer DK at least once significantly less often than the probability-based online panel respondents (17.6% versus 14.8%, $\chi^2(1)=16.6$, $p=0.00$, see DK bars in Figure 2).

However, looking at the proportions of DK across the online panels in detail (see left part of Figure 3), we find high variability in the proportion of people who chose DK at least once in the nonprobability online panels (between 9.1% in Panel 4 and 17.8% in Panel 7) as well as the probability-based online panels (between 15.4% in Panel 1 and 20.8% in GP). While we observe significantly more respondents in the GP and significantly fewer in Panel 4 answered DK at least once, the overall variability across the panels is so large that any differences between probability-based and nonprobability online panels seem coincidental.

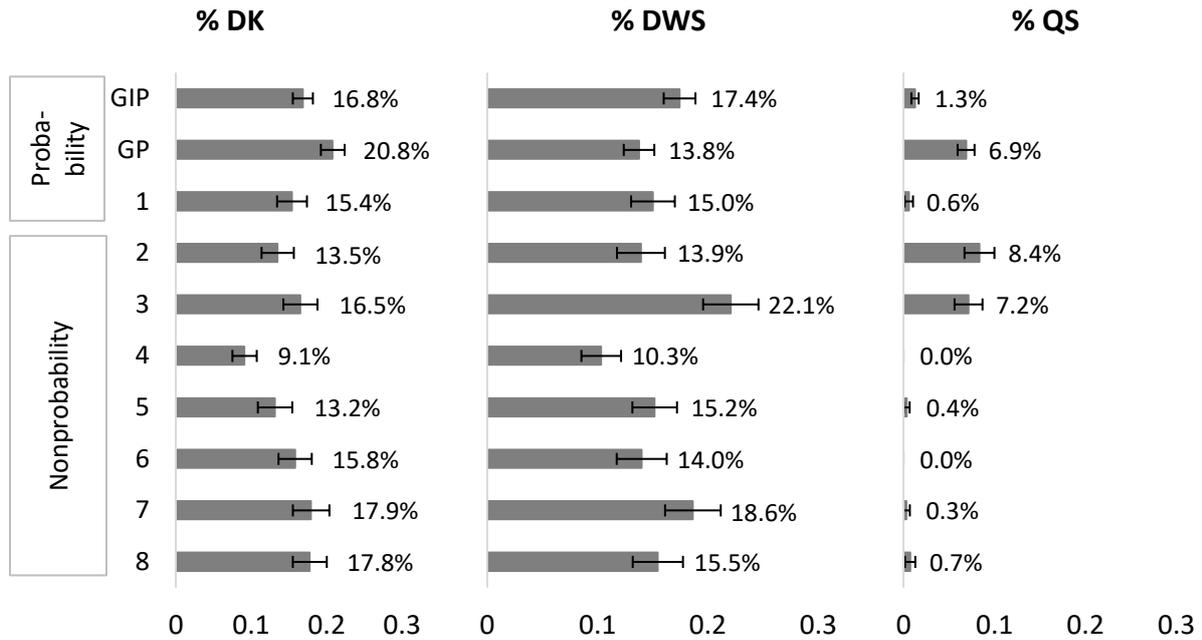


Figure 3: Proportion of DK, DWS, and QS across panels (bars); bootstrapped 95% confidence intervals (spikes)

Regarding our hypothesis that a higher proportion of respondents chooses to answer DWS in nonprobability online panels than in probability-based online panels (H2b), we find that there is no significant difference in the proportions of respondents that chose DWS at least once between the nonprobability online panels and the probability-based online panels (15.7% versus 15.5%, $\chi^2(1)=0.09$, $p>0.1$).

Examining the proportions of DWS across the online panels in detail (see middle part of Figure 3), we also do not find any generalizable evidence that may distinguish nonprobability online panels (between 10.3% in Panel 4 and 22.1% in Panel 3) from probability-based online panels (between 13.8% in GP and 17.4% in GIP).

With regard to our hypothesis that a higher proportion of respondents chooses to skip a question in nonprobability online panels than in probability-based online panels (H2c), we find no significant difference between nonprobability online panels and probability-based online panels (2.4% versus 2.8%, $\chi^2(1)=1.91$, $p>0.1$).

When we examine QS across panels in detail (see right part of Figure 3), we also find no generalizable pattern of differences between the nonprobability online panels (between 0.0% in Panel 4 and Panel

6 and 8.4% in Panel 2) and the probability-based online panels (between 0.6% in Panel 1 and 6.9% in GP).

Overall, we find that there is great variability across the online panels with regard to the amount and type of item nonresponse. This variability cannot be attributed to the sampling design of the panels because while Panel 3 has the highest proportion of general INR and also the highest proportion of respondents who answered DWS at least once, the GP has the highest proportion of respondents who answered DK at least once, and Panel 2 has the highest proportion of QS.

6.3 Midpoint selection

In Table 3, we present the results of our experiment on the influence of the answer scale design on the selection of the visual midpoint. Contrary to our hypothesis on midpoint selection (H3), we find no significant difference between nonprobability online panels and probability-based online panels in the effect of the visual design of the scale on midpoint selection. In the following, we investigate the differences in midpoint selection across nonprobability online panels and probability-based online panels in more detail based on our hypotheses 3a and 3b.

Table 3: Proportions and absolute numbers of respondents in the middle category by experimental group in probability-based online panels and nonprobability online panels

	<i>5-point scale</i>				<i>7-point scale</i>			
	Non-probability		Probability-based		Non-probability		Probability-based	
	%	<i>N</i>	%	<i>N</i>	%	<i>N</i>	%	<i>N</i>
<i>Hypothesis 3a</i>								
Condition 1: Conceptual midpoint matches visual midpoint	28.5	500	26.7	444	17.1	299	15.2	254
Condition 2: Conceptual midpoint but no visual midpoint	27.6	493	26.3	432	16.3	288	18.4	298
Difference between conditions 1 and 2	0.8	7	0.4	12	0.7	11	-3.2	-44
<i>Hypothesis 3b</i>								
Condition 3: Visual midpoint	14.7	254	9.2	154	10.4	181	11.7	196
Condition 4: No midpoint	14.0	246	11.9	194	11.6	205	10.6	175
Difference between conditions 3 and 4	0.7	8	-2.7	-40	-1.2	-24	1.1	21

With regard to our hypothesis that the difference between the proportion of respondents choosing the conceptual midpoint when it matches the visual midpoint (condition 1) and the proportion of respondents choosing the conceptual midpoint when there is no visual midpoint (condition 2) will be higher in nonprobability online panels than in probability-based online panels (H3a), we find that the difference in proportions is not significantly higher in the nonprobability online panels than in the probability-based online panels (0.8%-points and 0.4%-points respectively on the 5-point scale, $\chi^2(1)=0.01$, $p>0.1$; 0.7%-points and -3.2%-points respectively, $\chi^2(1)=-0.05$, $p>0.1$; on the 7-point scale). This indicates that respondents in nonprobability online panels are not more influenced by the visual design of the midpoint than respondents in probability-based online panels.

Regarding our hypothesis that the difference between the proportion of respondents choosing an answer option when it is located at the visual midpoint (condition 3) and the proportion of respondents choosing the same answer option when it is not located at the visual midpoint (condition 4) will be higher in nonprobability online panels than in probability-based online panels (H3b), we find that the difference in proportions is not significantly higher in the nonprobability online panels than in the probability-based online panels (0.7%-points and -2.7%-points respectively on the 5-point scale, $\chi^2=0.24$, $p>0.1$; -1.2%-points and 1.1%-points respectively, $\chi^2=0.05$, $p>0.1$; on the 7-point scale). In accordance with our findings from H3a, the findings on hypothesis H3b also indicate that respondents in nonprobability online panels are not more influenced by the visual design of the midpoint than respondents in probability-based online panels.

Furthermore, when examining the results of the experiment in each of the online panels (see Table C1 in Appendix C) we find that the visual design of the midpoint has no effect. Exceptions are three of the nonprobability online panels (Panel 5, Panel 6, and Panel 7), where we find a weakly significant effect in one comparison of experimental conditions each, but given the number of effects tested, this may well be just by chance.

6.4 Response quality and costs

Table A1 in Appendix A shows the total costs for data collection in the commercial online panels. Panel 1 participated without billing any costs. All other commercial online panels costed different amounts of money ranging from 5,392.97€ in Panel 2 to 10,676.44€ in Panel 8. There is no indication that the more costly panels perform better than the less costly panels in terms of data quality. For example, Panel 2 as the least costly commercial online panel has a slightly lower proportion of straight-lining, DK answers, and DWS answers than Panel 8 as the most costly commercial online panel. Panel 8, however, has a lower percentage of QS answers than Panel 2. Regarding the midpoint design experiment, both Panel 2 and Panel 8 do not show any significant differences across experimental subgroups. We therefore conclude that there is no association between costs and response quality.

7. Discussion

In this paper, we investigate the effect of respondent motivation on response quality in nonprobability and probability-based online panels. In our study, we implemented the same survey across ten online panels (seven nonprobability online panels and three probability-based online panels) during the same fieldwork period. In our analysis, we used three satisficing indicators: straight-lining in grid questions, item nonresponse, and midpoint selection in a visual design experiment. These indicators are associated with different degrees of satisficing. This research design provides us with a uniquely rich database for comparisons of response quality across online panels. It allows us to answer the question of whether the more incentive-oriented motivation of nonprobability online panel respondents negatively influences the quality of responses in comparison with respondents in probability-based online panels.

Overall, we find that response quality is worse in the nonprobability online panels than in the probability-based online panels with regard to straight-lining, but neither item nonresponse nor midpoint selection are differentially affected. In the following, we summarize and discuss the findings on each of our hypotheses in more detail.

In our study, we find evidence in support of our first hypothesis. Overall, there is significantly more straight-lining in the nonprobability online panels than on the probability-based online panels (13.4% versus 6.5%). When we examine the panels individually, we find that there is more straight-lining in all of the nonprobability online panels than there is in any of the probability-based online panels. These findings suggest that response quality is indeed lower in nonprobability online panels than in probability-based online panels, possibly due to a lower motivation to provide good answers in the former.

When analyzing item nonresponse, we find no generalizable evidence for our second hypothesis. Nonprobability online panels do not score higher on any of the item nonresponse indicators (DK, DWS, and QS) investigated. On all types of item nonresponse there is, however, a substantial amount of variability in response quality across the online panels in general.

In fact, contrary to our hypothesis H2a, we find that there are significantly less DK responses in nonprobability online panels than in probability-based online panels. This finding is, however, not consistent, but instead largely driven by one nonprobability panel (Panel 4) with a significantly lower proportion of DK answers than any other online panel, as well as one probability-based online panel (GP) with a significantly higher proportion of DK answers than all other online panels.

This suggests that other factors than the nonprobability versus probability-based characteristic might influence the proportion of DKs across online panels. Thus answering DK might be used as a satisficing strategy, but may be related to other characteristics than a respondent's motivation for participating in a panel. A potential explanation for the variability across online panels regarding DK responses might be question fatigue. In panels that use the questions from our survey (such as vote choice at the last general election) relatively often, respondents might defect from giving a substantive answer, while in panels that usually mainly conduct other types of surveys, for example on consumer surveys, respondents might still be willing to provide a substantive answer. Future

research will need to further investigate the reasons for the large variability in the proportion of DK answers across surveys.

With respect to DWS responses, we find similarly high variability across the online panels (between 10.3% in Panel 4 and 22.1% in Panel 3). Again, differences in DWS seem unrelated to the sampling design. In contrast, with respect to QS responses, we find considerably less variability. Many panels have either no QS at all or very little (between 0.3% and 1.3%). Only three panels have substantial amounts of QS (GP, Panel 2, and Panel 3). These differences in the proportion of QS found across panels may indicate that some online panels monitor question-skipping behavior and react to it in different ways (for example by excluding respondents that skip questions from the dataset or by probing respondents that skip questions), while other online panels do not take precautions against question-skipping. The GP, for instance, does not usually probe respondents who click on the “Next” button without responding to the question, while the default in GIP surveys is to softly probe respondents that try to skip a question.

We also find no evidence in support of our third hypothesis. The difference in the proportion of visual midpoint answers is not significantly higher in the nonprobability online panels than in the probability-based online panels. This might be because generally the differences in visual midpoint answers are small across experimental conditions and panels. Overall, we see little effect of the visual design of the midpoint on response quality.

Finally, we find that there is no association between response quality and survey costs across the nonprobability online panels. This means that investing in a more expensive panel does not automatically lead to better data.

Our study is one of very few that explore differences between nonprobability and probability-based online panels with regard to response quality. While we only detect differences in terms of straight-lining future studies will need to investigate whether our findings and null-findings replicate. Such replications will be particularly interesting with respect to the influence of the visual design of the

answer scale on midpoint selection reported in the literature (see for example Tourangeau et al. 2004), because using our experimental design we could not replicate such an effect, neither for the nonprobability nor for the probability-based online panels. We therefore hope that our study encourages further research into the topic of response quality in nonprobability and probability-based online panels.

Literature

- American Association for Public Opinion (2016). *Standard Definitions. Final Dispositions of Case Codes and Outcome Rates for Surveys*. Retrieved from https://www.aapor.org/AAPOR_Main/media/MainSiteFiles/NPS_TF_Report_Final_7_revised_FNL_6_22_13.pdf.
- Alwin, Duane F. and Jon A. Krosnick. 1991. "Aging, Cohorts, and the Stability of Sociopolitical Orientations Over the Life Span." *The American Journal of Sociology* 97(1):169–195.
- Baker, Reg, Stephen J. Blumberg, J. Michael Brick, Mick P. Couper, Melanie Courtright, J. Michael Dennis, Don Dillman, Martin R. Frankel, Philip Garland, Robert M. Groves, Courtney Kennedy, Jon Krosnick, Paul J. Lavrakas, Sunghee Lee, Michael Link, Linda Piekarski, Kumar Rao, Randell K. Thomas, and Dan Zahs. 2010. "Research Synthesis. AAPOR Report on Online Panels." *Public Opinion Quarterly* 74(4):711–781.
- Blom, Annelies G., Christina Gathmann, and Ulrich Krieger. 2015. "Setting up an online panel representative of the general population: The German Internet Panel." *Field methods* 27(4):391–408.
- Blom, Annelies G., Jessica M. Herzing, Carina Cornesse, Joseph W. Sakshaug, Ulrich Krieger, and Dayana Bossert. 2017. "Does the recruitment of offline households increase the sample representativeness of probability-based online panels? Evidence from the German Internet Panel." *Social Science Computer Review*, 35(4):498–520.
- Bosnjak, Michael, Tanja Dannwolf, Tobias Enderle, Ines Schaurer, Bella Struminskaya, Angela Tanner, and Kai W. Weyandt. 2017. "Establishing an Open Probability-Based Mixed-Mode Panel of the General Population in Germany." *Social Science Computer Review* 21(1):1–13.
- Bradburn, Norman M. and Seymour Sudman. 1988. *Polls & surveys: Understanding what they tell us*. San Francisco: Jossey-Bass.
- Cacioppo, John T. and Richard E. Petty. 1982. "The Need for Cognition." *Journal of Personality and Social Psychology* 42(1):116–131.

- Cacioppo, John T., Richard E. Petty, Jeffrey A. Feinstein, and W. Blair G. Jarvis. 1996. "Dispositional Differences in Cognitive Motivation: The Life and Times of Individuals Varying in Need for Cognition." *Psychological Bulletin* 119(2):197–253.
- Chang, Linchiat and Jon A. Krosnick. 2009. „National Surveys Via Rdd Telephone Interviewing Versus the Internet: Comparing Sample Representativeness and Response Quality." *Public Opinion Quarterly* 7(3):641–678.
- Converse, Philip E. 1974. "Nonattitudes and American Public Opinion: Comment: The Status of Nonattitudes." *American Political Science Review* 68(2):650–660.
- Couper, Mick P. 2008. *Designing Effective Web Surveys*. New York: Cambridge University Press.
- Couper, Mick P., Michael W. Traugott, and Mark J. Lamias. 2001. "Web Survey Design and Administration." *Public Opinion Quarterly* 65(2):230–253.
- Feick, Lawrence. 1989. "Latent class analysis of survey questions that include don't know responses." *Public Opinion Quarterly* 53(1):525–547.
- Fowler, Floyd J. and Charles F. Cannell. 1996. Using behavioral coding to identify cognitive problems with survey questions. In *Answering Questions: Methodology for Determining Cognitive and Communicative Processes in Survey Research* (pp.15–36). San Francisco: Jossey: Bass Publishers.
- Gilljam, Mikael and Donald Granberg. 1993. "Should We Take Don't Know for an Answer?" *Public Opinion Quarterly* 57(3):348–357.
- Goel, Sjarad, Adam Obeng, and David Rothschild. 2015. *Non-representative surveys: Fast, cheap, and mostly accurate*. Working paper, 27. Retrieved from <http://researchdmr.com/FastCheapAccurate>
- Green, Melanie C., Jon A. Krosnick, and Allyson L. Holbrook. 2001. *The Survey Response Process in Telephone and Face-to-Face Surveys: Differences in Respondent Satisficing and Social Desirability Response Bias*. Manuscript: Ohio State University.
- GreenBook. 2017. *GRIT CPR Report - 2017 Global Respondent Engagement Study*. Retrieved from https://www.greenbook.org/images/GRIT/2017_Q1/GRIT17Q1_2.pdf.

- Greszki, Robert, Marco Meyer, and Harald Schoen. 2014. The impact of speeding on data quality in nonprobability and freshly recruited probability-based online panels. In *Online Panel Research: A Data Quality Perspective* (pp.238–262). United Kingdom: John Wiley & Sons.
- Groves, Robert M., Stanley Presser, and Sarah Dipko. 2004. “The Role of Topic Interest in Survey Participation Decisions.” *Public Opinion Quarterly* 68(1):2–31.
- Hillygus, D. Sunshine, Natalie Jackson, and McKenzie Young. 2014. Professional respondents in non-probability online panels. In *Online Panel Research: A Data Quality Perspective* (pp. 219–237). United Kingdom: John Wiley & Sons.
- Holbrook, Allyson L., Melanie C. Green, and Jon A. Krosnick. 2003. “Telephone versus face-to-face interviewing of national probability samples with long questionnaires.” *Public Opinion Quarterly* 67(1):79–125.
- Kaminska, Olena, Allan L. McCutcheon, and Jaak Billiet. 2010. “Satisficing Among Reluctant Respondents in a Cross-National Context.” *Public Opinion Quarterly* 74(5):956–984.
- Kennedy, Courtney, Andrew Mercer, Scott Keeter, Nick Hatley, Kyley McGeeney, and Alejandra Gimenez. 2016. *Evaluating Online Nonprobability Surveys. Vendor choice matters; widespread errors found for estimates based on blacks and Hispanics*. Pew Research Center. Retrieve from <http://www.pewresearch.org/files/2016/04/Nonprobability-report-May-2016-FINAL.pdf>.
- Keusch, Florian, Bernad Batinic, and Wolfgang Mayerhofer. 2014. Motives for joining nonprobability online panels and their association with survey participation behavior. In *Online Panel Research: A Data Quality Perspective* (pp.171–191). United Kingdom: John Wiley & Sons.
- Klocklars, Alan J. and Midori Yamagashi. 1988. “The Influence of Labels and Positions in Rating Scales.” *Journal of Educational Measurement* 25(2):85–96.
- Knäuper, Bärbel. 1999. “The Impact of Age and Education on Response Order Effects in Attitude Measurement.” *Public Opinion Quarterly* 63(3):347–370.

- Krosnick, Jon A. 1991. "Response Strategies for Coping with the Cognitive Demands of Attitude Measures in Surveys." *Applied Cognitive Psychology* 5(3):213–236.
- Krosnick, Jon A. 1992. The Impact of Cognitive Sophistication and Attitude Importance on Response-Order and Question-Order Effects. In *Context Effects in Social and Psychological Research* (pp.203–218). NewYork: Springer Verlag.
- Krosnick, Jon A. 1999. Maximizing Questionnaire Quality. In *Measures of Political Attitudes* (pp. 37–57). Oxford: Elsevier LTD.
- Krosnick, Jon A. and Duane F. Alwin. 1987. "An evaluation of a cognitive theory of response-order effects in survey measurement." *Public Opinion Quarterly* 51(2):201–219.
- Krosnick, Jon A. and Matthew K. Berent. 1993. "Comparisons of Party Identification and Policy Preferences: The Impact of Survey Question Format." *American Journal of Political Science* 73(3):941–964.
- Krosnick, Jon A. and Leandre R. Fabrigar. 1997. Designing rating scales for effective measurement in surveys. In *Survey measurement and process quality* (pp.141-164). USA: John Wiley & Sons, Inc.
- Krosnick, Jon A., Allyson L. Holbrook, Matthew K. Berent, Richard T. Carson, W. Michael Hanemann, Raymond J. Kopp, Robert Cameron Mitchell, Stanley Presser, Paul A. Ruud, V. Kerry Smith, Wendy R. Moody, Melanie C. Green, and Michael Conaway. 2002. "The Impact of "No Opinion" Response Options on Data Quality." *Public Opinion Quarterly* 66(1):371–403.
- Legleye, Stéphane, Géraldine Charrance, Nicolas Razafindratsima, Nathalie Bajos, Aline Bohet, Caroline Moreau, and the FECOND research Team. 2015. The Use of a Nonprobability Internet Panel to Monitor Sexual Reproductive Health in the General Population. *Sociological Methods & Research* 1-35.
- Loosveldt, Geert and Nathalie Sonck. 2008. „An evaluation of the weighting procedures for an online access panel survey." *Survey research methods* 2(2):93–105.

- Malhotra, Neil. 2008. "Completion Time and Response Order Effects in Web Surveys." *Public Opinion Quarterly* 72(5):914–934.
- Malhotra, Neil and Krosnick, Jon A. 2007. "The effect of survey mode and sampling on inferences about political attitudes and behavior: Comparing the 2000 and 2004 ANES to internet surveys with nonprobability samples." *Political Analysis* 15(3):286–323.
- Malhotra, Neil, Joanne M. Miller, and Justin Wedeking. 2014. The relationship between nonresponse strategies and measurement error. Comparing online panel surveys to traditional surveys. In *Online Panel Research: A Data Quality Perspective* (pp. 313–336). United Kingdom: John Wiley & Sons.
- Narayan, Sowmya and Jon A. Krosnick. 1996. "Education Moderates Some Response Effects in Attitude Measurement." *Public Opinion Quarterly* 60(1):58. <https://doi.org/10.1086/297739>
- Pasek, Josh. 2016. "When will Nonprobability Surveys Mirror Probability Surveys? Considering Types of Inference and Weighting Strategies as Criteria for Correspondence." *International Journal of Public Opinion Research* 28(2):269–291.
- Pennay, Darren Walter, Dina Neiger, Paul J. Lavrakas, Kim A. Borg, Sebastian Misson, and Nikki Honey. 2016, May. 2015-2016 Australian Online Panels Benchmarking Study: A Comparison of Surveys Using Probability and Nonprobability Samples in an Australian Research Context: Paper presented at the 69th World Association for Public Opinion Research (WAPOR) conference. Austin, Texas.
- Schonlau, Matthias and Vera Toepoel. 2015. "Straightlining in Web survey panels over time". *Survey Research Methods* 9(2):125–137. <https://doi.org/10.18148/srm/2015.v9i2.6128>
- Schuman, Howard and Stanley Presser. 1981. *Questions and Answers in Attitude Surveys*. New York: Academic Press.
- Smith, Tom W. 2003. "An Experimental Comparison of Knowledge Networks and the GSS." *International Journal of Public Opinion Research* 15(2):167–179.

- Sparrow, Nick. 2006. Developing reliable online polls. *International Journal of Market Research* 48(6):659.
- Stoop, Ineke A. 2005. *The hunt for the last respondent: Nonresponse in sample surveys*. Netherlands: Social and Cultural Planning Office.
- Sturgis, Patrick, Caroline Roberts, and Patten Smith. 2014. "Middle Alternatives Revisited." *Sociological Methods & Research* 43(1):15–38.
- Toepoel, Vera, Corrie Vis, Marcel Das, Arthur van Soest. 2009. Design of Web Questionnaires. An Information-Processing Perspective for the Effect of response Categories. *Sociological Methods & Research* 37(3): 371-392.
- Tourangeau, Roger, Mick P. Couper, and Frederick Conrad. 2004. "Spacing, Position, and Order Interpretive Heuristics for Visual Features of Survey Questions." *Public Opinion Quarterly* 68(3):368–393.
- Tourangeau, Roger, Lance J. Rips, and Kenneth Rasinski. 2000. *The psychology of survey response*. Cambridge, New York: Cambridge University Press.
- Wang, Wei, David Rothschild, Sharad Goel, and Andrew Gelman. 2015. „Forecasting elections with non-representative polls." *International Journal of Forecasting* 31(1):980–991.
- Yeager, David S., Jon A. Krosnick, LinChiat Chang, Harold S. Lavitz, Matthew S. Levendusky, Alberto Simpser, and Rui Wang. 2011. "Comparing the accuracy of RDD telephone surveys and internet surveys conducted with probability and non-probability samples." *Public Opinion Quarterly* 75(4):709–747.

Appendix A: Panel costs

Table A1: Total costs for the commercial online panels (final costs paid for three waves of fieldwork including 19 percent VAT; costs for GIP, and GP samples not available, because the samples were not specifically contracted for this study; instead, the data are freely available as Scientific Use Files from their respective data archives)

Panel	Total Costs (€)
1	Participated free of charge
2	5,392.97
3	5,618.57
4	7,061.11
5	7,411.00
6	7,636.22
7	8,380.46
8	10,676.44

Appendix B: Question texts and answer scales by indicator

Variable	German	English
Environmental zones	<p>Fragetext Wie ist Ihre Meinung zu Umweltzonen in Großstädten?</p> <p>Antworten Ich finde Umweltzonen in Großstädten ...</p> <ul style="list-style-type: none"> • sehr gut • gut • eher gut • mittelmäßig • eher schlecht • schlecht • sehr schlecht • weiß nicht <p>Fehlermeldung [WENN gar keine Angabe gemacht wurde:] Sie haben noch keine Antwort gegeben. Bitte suchen Sie die entsprechende Antwort aus. Falls Sie keine Angabe machen möchten, klicken Sie bitte auf *Weiter*.</p>	<p>Question text What is your opinion regarding environmental zones in large cities?</p> <p>Answer options I think environmental zones in large cities are...</p> <ul style="list-style-type: none"> • Very good • Good • Somewhat good • Neither good nor bad • Somewhat bad • Bad • Very bad • Don't know <p>Error message [IF no answer was given:] You did not yet give an answer. Please select the respective answer. If you do not want to give an answer, click on *Next*.</p>
Driver's licence (yes/no)	<p>Fragetext Wurde Ihnen in Deutschland jemals ein Führerschein ausgestellt? Damit ist jeder in Deutschland gültige Führerschein beziehungsweise jede Fahrerlaubnis gemeint – unabhängig von der Fahrzeugklasse. Bitte wählen Sie auch „ja“ aus, wenn Sie Ihren Führerschein (vorübergehend) abgegeben haben.</p> <p>Antworten • ja • nein</p> <p>Fehlermeldung [WENN gar keine Angabe gemacht wurde:] Sie haben noch keine Antwort gegeben. Bitte suchen Sie die entsprechende Antwort aus. Falls Sie keine Angabe machen möchten,</p>	<p>Question text Did you ever receive a driver's license in Germany? By this we mean any driver's license that is valid in Germany -- irrespective of vehicle classes. Please also select "yes" if you have (temporarily) handed in your driver's license.</p> <p>Answer options • yes • no</p> <p>Error message [IF no answer was given:] You did not yet give an answer. Please select the respective answer. If you do not want to give an</p>

	klicken Sie bitte auf *Weiter*.	answer click on *Next*.
Driver's licence (since 1999)	<p>Fragetext Wurde Ihnen nach dem 1. Januar 1999 ein Führerschein ausgestellt?</p> <p>Hinweistext Das kann zum Beispiel sein, weil Sie... ... einen neuen Führerschein gemacht haben, ... Ihren Führerschein verloren hatten und er Ihnen neu ausgestellt wurde oder ... Ihren Führerschein eingetauscht haben. Bei Führerscheinen, die in Deutschland nach dem 1. Januar 1999 neu ausgestellt wurden, handelt es sich um Kartenführerscheine, wie diesen:</p>  <p>Quelle: Bundesdruckerei GmbH</p> <p>Antworten</p> <ul style="list-style-type: none"> • ja • nein <p>Fehlermeldung [WENN gar keine Angabe gemacht wurde:] Sie haben noch keine Antwort gegeben. Bitte suchen Sie die entsprechende Antwort aus. Falls Sie keine Angabe machen möchten, klicken Sie bitte auf *Weiter*.</p>	<p>Question text Did you receive a driver's license after January 1st, 1999?</p> <p>Help text This could be because you... ... got a new driver's license, ... lost your driver's license and received a new one, ... swapped your driver's license. Driver's licenses that were issued after January 1st, 1999 are card driver's licenses that look like this:</p>  <p>Source: Bundesdruckerei GmbH</p> <p>Answer options</p> <ul style="list-style-type: none"> • yes • no <p>Error message [IF no answer was given:] You did not yet give an answer. Please select the respective answer. If you do not want to give an answer click on *Next*.</p>
Driver's licence (foreign)	<p>Fragetext Hatten Sie jemals einen ausländischen Führerschein, mit dem Sie in Deutschland ein Auto fahren durften?</p>	<p>Question text Did you ever have a foreign driver's license that allowed you to drive a car in Germany?</p>

	<p>Hilfetext Bitte wählen Sie auch „ja“ aus, wenn Sie Ihren Führerschein (vorübergehend) abgeben mussten.</p> <p>Antworten</p> <ul style="list-style-type: none"> • ja • nein <p>Fehlermeldung [WENN gar keine Angabe gemacht wurde:] Sie haben noch keine Antwort gegeben. Bitte suchen Sie die entsprechende Antwort aus. Falls Sie keine Angabe machen möchten, klicken Sie bitte auf *Weiter*.</p>	<p>Help text Please also select "yes" if you have (temporarily) handed in your driver's license.</p> <p>Answer options</p> <ul style="list-style-type: none"> • yes • no <p>Error message [IF no answer was given:] You did not yet give an answer. Please select the respective answer. If you do not want to give an answer click on *Next*.</p>																																																																																								
Traffic violation points	<p>Fragetext Hatten Sie am 1. Januar 2015 Punkte für Verkehrsverstöße, also sogenannte Punkte in Flensburg? Und wenn ja, wie viele?</p> <p>Hilfetext Bitte geben Sie Ihre Punkte nach dem neuen System (gültig ab dem 1. Mai 2014) an, Informationen dazu finden Sie in der Abbildung. Denken Sie auch daran, dass Punkte nach einer bestimmten Zeit wieder gelöscht werden. Geben Sie daher bitte nur die Punkte an, die Sie am 1. Januar 2015 hatten.</p> <table border="1" data-bbox="297 951 687 1278"> <thead> <tr> <th colspan="4">Überführung der Punktestände</th> </tr> <tr> <th colspan="4">Verkehrszentralregister (VZR) vs. Fahreignungsregister (FAER)</th> </tr> <tr> <th>Punktestand am 30.04.2014</th> <th></th> <th>Zuordnung im Fahreignungs-Bewertungssystem am 01.05.2014</th> <th></th> </tr> </thead> <tbody> <tr> <td>1-3</td> <td>→</td> <td>1</td> <td></td> </tr> <tr> <td>4-5</td> <td>→</td> <td>2</td> <td>Vormerkung</td> </tr> <tr> <td>6-7</td> <td>→</td> <td>3</td> <td></td> </tr> <tr> <td>8-10</td> <td>→</td> <td>4</td> <td></td> </tr> <tr> <td>11-13</td> <td>→</td> <td>5</td> <td>Ermahnung</td> </tr> <tr> <td>14-15</td> <td>→</td> <td>6</td> <td></td> </tr> <tr> <td>16-17</td> <td>→</td> <td>7</td> <td>Verwarnung</td> </tr> <tr> <td>≥18</td> <td>→</td> <td>8</td> <td>Entziehung</td> </tr> </tbody> </table> <p><small>Quelle: Bundesministerium für Verkehr und digitale Infrastruktur</small></p> <p>Bei dieser Frage können Sie nur eine Antwort geben.</p> <p>Antworten</p>	Überführung der Punktestände				Verkehrszentralregister (VZR) vs. Fahreignungsregister (FAER)				Punktestand am 30.04.2014		Zuordnung im Fahreignungs-Bewertungssystem am 01.05.2014		1-3	→	1		4-5	→	2	Vormerkung	6-7	→	3		8-10	→	4		11-13	→	5	Ermahnung	14-15	→	6		16-17	→	7	Verwarnung	≥18	→	8	Entziehung	<p>Question text Did you have traffic violation points on January 1st, 2015, that is to say the so-called "Punkte in Flensburg"? And if you did, how many did you have?</p> <p>Help text Please report the points in compliance with the new system (valid from May 1, 2014). You can find information on this in the figure below. Remember that points are deleted after a certain time. Please therefore only report points that you had on January 1st, 2015.</p> <table border="1" data-bbox="1216 951 1606 1278"> <thead> <tr> <th colspan="4">Überführung der Punktestände</th> </tr> <tr> <th colspan="4">Verkehrszentralregister (VZR) vs. Fahreignungsregister (FAER)</th> </tr> <tr> <th>Punktestand am 30.04.2014</th> <th></th> <th>Zuordnung im Fahreignungs-Bewertungssystem am 01.05.2014</th> <th></th> </tr> </thead> <tbody> <tr> <td>1-3</td> <td>→</td> <td>1</td> <td></td> </tr> <tr> <td>4-5</td> <td>→</td> <td>2</td> <td>Vormerkung</td> </tr> <tr> <td>6-7</td> <td>→</td> <td>3</td> <td></td> </tr> <tr> <td>8-10</td> <td>→</td> <td>4</td> <td></td> </tr> <tr> <td>11-13</td> <td>→</td> <td>5</td> <td>Ermahnung</td> </tr> <tr> <td>14-15</td> <td>→</td> <td>6</td> <td></td> </tr> <tr> <td>16-17</td> <td>→</td> <td>7</td> <td>Verwarnung</td> </tr> <tr> <td>≥18</td> <td>→</td> <td>8</td> <td>Entziehung</td> </tr> </tbody> </table> <p><small>Quelle: Bundesministerium für Verkehr und digitale Infrastruktur</small></p> <p>You can only select one answer to this question.</p> <p>Answer options</p>	Überführung der Punktestände				Verkehrszentralregister (VZR) vs. Fahreignungsregister (FAER)				Punktestand am 30.04.2014		Zuordnung im Fahreignungs-Bewertungssystem am 01.05.2014		1-3	→	1		4-5	→	2	Vormerkung	6-7	→	3		8-10	→	4		11-13	→	5	Ermahnung	14-15	→	6		16-17	→	7	Verwarnung	≥18	→	8	Entziehung
Überführung der Punktestände																																																																																										
Verkehrszentralregister (VZR) vs. Fahreignungsregister (FAER)																																																																																										
Punktestand am 30.04.2014		Zuordnung im Fahreignungs-Bewertungssystem am 01.05.2014																																																																																								
1-3	→	1																																																																																								
4-5	→	2	Vormerkung																																																																																							
6-7	→	3																																																																																								
8-10	→	4																																																																																								
11-13	→	5	Ermahnung																																																																																							
14-15	→	6																																																																																								
16-17	→	7	Verwarnung																																																																																							
≥18	→	8	Entziehung																																																																																							
Überführung der Punktestände																																																																																										
Verkehrszentralregister (VZR) vs. Fahreignungsregister (FAER)																																																																																										
Punktestand am 30.04.2014		Zuordnung im Fahreignungs-Bewertungssystem am 01.05.2014																																																																																								
1-3	→	1																																																																																								
4-5	→	2	Vormerkung																																																																																							
6-7	→	3																																																																																								
8-10	→	4																																																																																								
11-13	→	5	Ermahnung																																																																																							
14-15	→	6																																																																																								
16-17	→	7	Verwarnung																																																																																							
≥18	→	8	Entziehung																																																																																							

	<ul style="list-style-type: none"> • 0 Punkte • 1 Punkt • 2 Punkte • 3 Punkte • 4 Punkte • 5 Punkte • 6 Punkte • 7 Punkte • 8 Punkte • weiß nicht <p>Fehlermeldung [WENN gar keine Angabe gemacht wurde:] Sie haben noch keine Antwort gegeben. Bitte suchen Sie die entsprechende Antwort aus. Falls Sie keine Angabe machen möchten, klicken Sie bitte auf *Weiter*.</p>	<ul style="list-style-type: none"> • 0 points • 1 point • 2 points • 3 points • 4 points • 5 points • 6 points • 7 points • 8 points • Don't know <p>Error message [IF no answer was given:] You did not yet give any answer. Please select the respective answer. If you do not want to give an answer click on *Next*.</p>
Number of cars	<p>Fragetext Waren am 1. Januar 2015 auf Sie persönlich Autos zugelassen? Und wenn ja, wie viele? Damit meinen wir PKWs, bei denen Sie persönlich im Fahrzeugschein als Halter eingetragen waren. Bitte beziehen Sie dabei ausschließlich PKWs ein: Das sind Fahrzeuge zur Beförderung von Personen mit mindestens vier Rädern und mit höchstens acht Sitzplätzen außer dem Fahrersitz. Wohnmobile zählen zu PKWs, nicht eingeschlossen sind jedoch Wohnmobile auf LKW-Basis.</p> <p>Antworten</p> <ul style="list-style-type: none"> • Nein, es war kein PKW auf mich persönlich zugelassen. • Ja, auf mich persönlich waren PKWs zugelassen. Anzahl PKW: [offenes Feld] <p>Fehlermeldungen [WENN gar keine Angabe gemacht wurde:] Sie haben noch keine Antwort gegeben. Bitte suchen Sie die entsprechende Antwort aus. Falls Sie keine Angabe machen möchten, klicken Sie bitte auf *Weiter*.</p>	<p>Question text Were there any cars registered in your name on January 1st, 2015? And if there were, how many were there? With this we mean automobiles for which you personally are registered as the owner in the vehicle registration certificate. Please exclusively take into account only automobiles: That is vehicles for transportation of persons with at least four wheels and a maximum of eight seats apart from the driver's seat. Recreational vehicles are automobiles, but recreational vehicles based on trucks are not included.</p> <p>Answer options</p> <ul style="list-style-type: none"> • No, there weren't any cars registered in my name. • Yes, there were cars registered in my name. Number of cars: [open answer box] <p>Error messages [IF no answer was given:] You did not yet give an answer. Please select the respective answer. If you do not want to give an answer click on *Next*.[IF answer</p>

	<p>[WENN Antwort 2 „Ja, auf mich persönlich waren PKWs zugelassen. Anzahl PKW:“ angeklickt, aber keine Eingabe gemacht wurde] Bitte tragen Sie Ihre Antwort in das Feld ein. [WENN Eingabe in offenem Feld gemacht, aber andere Antwort angeklickt wurde] Sie haben „Nein, es war kein PKW auf mich persönlich zugelassen.“ ausgewählt und gleichzeitig eine Anzahl eingetragen. Bitte löschen Sie entweder diese Anzahl oder wählen Sie „Ja, auf mich persönlich waren PKWs zugelassen. Anzahl PKW:“ aus, damit Ihre Antwort eindeutig ist. [WENN keine ganze Zahl eingegeben wurde] Bitte tragen Sie eine ganze Zahl zwischen 0 und 99 ein.</p>	<p>option 2 „Yes, there were cars registered in my name. Number of cars:“ was selected, but no entry was made:] Please enter an answer into the box. [IF an entry was made in the box but another answer selected:] You have selected „No, there weren't any cars registered in my name. “ and at the same time entered a number. Please either delete the number or select „Yes, there were cars registered in my name. Number of cars:“ to provide a distinctive answer. [IF no number was entered] Please enter a number between 0 and 99.</p>
Health	<p>Fragetext Was würden Sie sagen, wie ist Ihre Gesundheit im Allgemeinen? Antworten</p> <ul style="list-style-type: none"> • sehr gut • gut • mittelmäßig • schlecht • sehr schlecht • weiß nicht <p>Fehlermeldung [WENN gar keine Angabe gemacht wurde:] Sie haben noch keine Antwort gegeben. Bitte suchen Sie die entsprechende Antwort aus. Falls Sie keine Angabe machen möchten, klicken Sie bitte auf *Weiter*.</p>	<p>Question text What would you say, how is your health generally? Answer options</p> <ul style="list-style-type: none"> • Very good • Good • Neither good nor bad • Bad • Very bad • Don't know <p>Error message [IF no answer was given:] You did not yet give an answer. Please select the respective answer. If you do not want to give an answer click on *Next*.</p>
Health insurance	<p>Fragetext Sind Sie krankenversichert? Hilfetext Private Zusatzversicherungen für zusätzliche Leistungen sind nicht gemeint. Antworten Ja, und zwar in einer gesetzlichen Krankenversicherung</p>	<p>Question text Do you have any health insurance? Help text Additional private supplementary insurances are not meant. Answer options Yes, namely in a public health insurance</p> <ul style="list-style-type: none"> • own mandatory insurance

	<ul style="list-style-type: none"> • selbst pflichtversichert • selbst freiwillig versichert • als Familienangehörige/-r versichert ... in einer privaten Krankenversicherung • selbst versichert • als Familienangehörige/-r versichert • Nein, nicht krankenversichert <p>Fehlermeldung [WENN gar keine Angabe gemacht wurde:] Sie haben noch keine Antwort gegeben. Bitte suchen Sie die entsprechende Antwort aus. Falls Sie keine Angabe machen möchten, klicken Sie bitte auf *Weiter*.</p>	<ul style="list-style-type: none"> • own voluntary insurance • insured as a family member ... in a private health insurance • own insurance • insured as a family member • No, not health insured <p>Error message [IF no answer was given:] You did not yet give any answer. Please select the respective answer. If you do not want to give an answer click on *Next*.</p>
Height	<p>Frage Wie groß sind Sie?</p> <p>Antwort</p> <ul style="list-style-type: none"> • Geben Sie bitte Ihre Größe in Zentimetern (cm) an: [offenes Feld] • Keine Angabe <p>Fehlermeldungen [WENN gar keine Angabe gemacht wurde:] Sie haben noch keine Antwort gegeben. Bitte suchen Sie die entsprechende Antwort aus. Falls Sie keine Angabe machen möchten, klicken Sie bitte auf *Weiter*. [WENN Antwort „Geben Sie Ihre Größe in Zentimetern (cm) an:“ angeklickt, aber keine Eingabe gemacht wurde] Bitte tragen Sie Ihre Antwort in das Feld ein. [WENN Eingabe in offenem Feld gemacht, aber gleichzeitig „Keine Angabe“ angeklickt wurde] Sie haben „Keine Angabe“ ausgewählt und gleichzeitig eine Größe eingetragen. Bitte löschen Sie entweder die Größe oder wählen Sie „Geben Sie bitte Ihre Größe in Zentimetern (cm) an:“ aus, damit Ihre Antwort eindeutig ist. [WENN keine ganze Zahl eingegeben wurde] Bitte tragen Sie eine ganze Zahl zwischen 0 und 999 ein.</p>	<p>Question text How tall are you?</p> <p>Answer options</p> <ul style="list-style-type: none"> • Select height in centimeters (cm): [open box] • No response <p>Error messages [IF no answer was given:] You did not yet give an answer. Please select the respective answer. If you do not want to give an answer click on *Next*.[IF answer „Select height in centimeters (cm):“ was selected, but no entry was made:] Please enter an answer into the box. [IF an entry was made in the box but at the same time "No response" was selected:] You selected "No response" and entered a height at the same time. Please either delete the height or select "Select height in centimeters (cm):" for the answer to be distinct. [IF no number was entered:] Please enter a number between 0 and 999.</p>

Weight	<p>Fragetext Wie viel wiegen Sie?</p> <p>Antwort</p> <ul style="list-style-type: none"> • Geben Sie bitte Ihr Gewicht in Kilogramm (kg) an: [offenes Feld] • Keine Angabe <p>Fehlermeldungen</p> <p>[WENN gar keine Angabe gemacht wurde:] Sie haben noch keine Antwort gegeben. Bitte suchen Sie die entsprechende Antwort aus. Falls Sie keine Angabe machen möchten, klicken Sie bitte auf *Weiter*.</p> <p>[WENN Antwort „Geben Sie bitte Ihr Gewicht in Kilogramm (kg) an:“ angeklickt, aber keine Eingabe gemacht wurde] Bitte tragen Sie Ihre Antwort in das Feld ein.</p> <p>[WENN Eingabe in offenem Feld gemacht, aber gleichzeitig „Keine Angabe“ angeklickt wurde] Sie haben „Keine Angabe“ ausgewählt und gleichzeitig ein Gewicht eingetragen. Bitte löschen Sie entweder das Gewicht oder wählen Sie „Geben Sie bitte Ihr Gewicht in Kilogramm (kg) an:“ aus, damit Ihre Antwort eindeutig ist.</p> <p>[WENN keine ganze Zahl eingegeben wurde] Bitte tragen Sie eine ganze Zahl zwischen 0 und 999 ein.</p>	<p>Question text How much do you weigh?</p> <p>Answer options</p> <ul style="list-style-type: none"> • Please select a weight in kilogram (kg): [open box] • No response <p>Error messages</p> <p>[IF no answer was given:] You did not yet give an answer. Please select the respective answer. If you do not want to give an answer click on *Next*.</p> <p>[IF answer „Please select a weight in kilogram (kg):“ was selected, but no entry was made:] Please enter an answer into the box.</p> <p>[IF an entry was made into the box but at the same time "No response" was selected:] You selected "No response" and entered a height at the same time. Please either delete the weight or select "Please select a weight in kilogram (kg):" for the answer to be distinct.</p> <p>[IF no number was entered:] Please enter a number between 0 and 999.</p>
Voter's recall	<p>Fragetext Welche Partei haben Sie bei der letzten Bundestagswahl am 22. September 2013 mit Ihrer Zweitstimme gewählt?</p> <p>Hilfetext Bei dieser Frage können Sie nur eine Antwort geben.</p> <p>Antworten</p> <ul style="list-style-type: none"> • CDU/CSU • SPD • Die Linke • Bündnis 90/Die Grünen • FDP • Piratenpartei 	<p>Question text What party did you vote for in the last general election on September 22nd, 2013?</p> <p>Help text For this question, you can only select one answer.</p> <p>Answer options</p> <ul style="list-style-type: none"> • The Christian Democratic Union/Christian Social Union "CDU/CSU" • The Social Democratic Party "SPD" • The Left "Die Linke" • The Greens "Bündnis 90/Die Grünen" • The Free Democratic Party "FDP" • Pirate Party "Piratenpartei"

	<ul style="list-style-type: none"> • AfD (Alternative für Deutschland) • NPD • eine andere Partei, und zwar: [offenes Feld] • Ich war nicht wahlberechtigt. • Ich habe nicht gewählt. • Ich weiß es nicht mehr. • Ich möchte es nicht sagen. <p>Fehlermeldungen</p> <p>[WENN gar keine Angabe gemacht wurde:] Sie haben noch keine Antwort gegeben. Bitte suchen Sie die entsprechende Antwort aus. Falls Sie keine Angabe machen möchten, klicken Sie bitte auf *Weiter*.</p> <p>[WENN Antwort 9 „eine andere Partei, und zwar“ angeklickt, aber keine Eingabe gemacht wurde] Bitte tragen Sie Ihre Antwort in das Feld ein.</p> <p>[WENN Eingabe in offenem Feld gemacht, aber andere Antwort angeklickt wurde] Sie haben einen Text in das Feld hinter der Antwort „eine andere Partei, und zwar:“ eingetragen und gleichzeitig eine andere Antwort ausgewählt. Bitte löschen Sie entweder den Text in diesem Feld oder wählen Sie die Antwort „eine andere Partei, und zwar:“ aus, damit Ihre Antwort eindeutig ist.</p>	<ul style="list-style-type: none"> • Alternative for Germany "AfD (Alternative für Deutschland)" • National Democratic Party of Germany "NPD" • Other party, please enter: [open box] • Not eligible to vote. • I didn't vote. • Don't know. • Don't want to tell. <p>Error messages</p> <p>[IF no answer was given:] You did not yet give an answer. Please select the respective answer. If you do not want to give an answer click on *Next*.</p> <p>[IF answer 9 „Other party, please enter:“ was selected, but no entry was made:] Please enter an answer into the box.</p> <p>[IF an entry was made but no answer was selected:] You have entered text into the box after "Other party, please enter:" and at the same time selected an answer option. Please either delete the text in the box or choose the answer option "Other party, please enter:" to provide a distinct response.</p>
Party membership	<p>Fragetext Sind Sie Mitglied in einer Partei? Und wenn ja, in welcher?</p> <p>Hilfetext Bei dieser Frage können Sie nur eine Antwort geben.</p> <p>Antworten</p> <ul style="list-style-type: none"> • CDU/CSU • SPD • Die Linke • Bündnis 90/Die Grünen • FDP • Piratenpartei 	<p>Question text Are you a member of a political party? And if you are, which one?</p> <p>Help text For this question you can only select one answer.</p> <p>Answer options</p> <ul style="list-style-type: none"> • The Christian Democratic Union/Christian Social Union "CDU/CSU" • The Social Democratic Party "SPD" • The Left "Die Linke" • The Greens "Bündnis 90/Die Grünen" • The Free Democratic Party "FDP" • Pirate Party "Piratenpartei"

	<ul style="list-style-type: none"> • AfD (Alternative für Deutschland) • NPD • eine andere Partei, und zwar: [offenes Feld] • Ich bin in keiner Partei Mitglied. • Ich bin ausschließlich in der Jugendorganisation einer Partei Mitglied. • Ich weiß es nicht. • Ich möchte es nicht sagen. <p>Fehlermeldungen</p> <p>[WENN gar keine Angabe gemacht wurde:] Sie haben noch keine Antwort gegeben. Bitte suchen Sie die entsprechende Antwort aus. Falls Sie keine Angabe machen möchten, klicken Sie bitte auf *Weiter*.</p> <p>[WENN Antwort 9 „eine andere Partei, und zwar“ angeklickt, aber keine Eingabe gemacht wurde] Bitte tragen Sie Ihre Antwort in das Feld ein.</p> <p>[WENN Eingabe in offenem Feld gemacht, aber andere Antwort angeklickt wurde] Sie haben einen Text in das Feld hinter der Antwort „eine andere Partei, und zwar:“ eingetragen und gleichzeitig eine andere Antwort ausgewählt. Bitte löschen Sie entweder den Text in diesem Feld oder wählen Sie die Antwort „eine andere Partei, und zwar:“ aus, damit Ihre Antwort eindeutig ist.</p>	<ul style="list-style-type: none"> • Alternative for Germany "AfD (Alternative für Deutschland)" • National Democratic Party of Germany "NPD" • Other party, please enter: [open box] • I am not a party member. • I am exclusively the member of a party youth organization. • Don't know. • Don't want to say. <p>Error messages</p> <p>[IF no answer was given:] You did not yet give an answer. Please select the respective answer. If you do not want to give an answer click on *Next*.</p> <p>[IF answer option 9 „Other party, please enter: “ was selected but no entry was made:] Please enter an answer into the box.</p> <p>[IF an entry was made but no answer was selected:] You have entered text into the box after "Other party, please enter:" and at the same time selected an answer option. Please either delete the text in the box or choose the answer option "Other party, please enter:" to provide a distinct response.</p>
Political interest	<p>Fragetext Wie stark interessieren Sie sich im Allgemeinen für politische Themen?</p> <p>Antworten</p> <ul style="list-style-type: none"> • überhaupt nicht • • • • • • • sehr stark 	<p>Question text How strongly are you generally interested in political topics?</p> <p>Answer options</p> <ul style="list-style-type: none"> • not at all • • • • • • • very strongly

	<p>Fehlermeldung [WENN gar keine Angabe gemacht wurde:] Sie haben noch keine Antwort gegeben. Bitte suchen Sie die entsprechende Antwort aus. Falls Sie keine Angabe machen möchten, klicken Sie bitte auf *Weiter*.</p>	<p>Error message [IF no answer was given:] You did not yet give an answer. Please select the respective answer. If you do not want to give an answer click on *Next*.</p>
Big 5	<p>Fragetext Nun kommen einige allgemeine Aussagen, die zur Beschreibung von Personen verwendet werden können. Diese Aussagen können auf Sie persönlich mehr oder weniger zutreffen. Bitte geben Sie bei jeder Aussage an, inwieweit die Aussage auf Sie selbst zutrifft.</p> <ul style="list-style-type: none"> • Ich schenke anderen leicht Vertrauen, glaube an das Gute im Menschen. • Ich habe nur wenig künstlerisches Interesse. • Ich neige dazu, andere zu kritisieren. • Ich habe eine aktive Vorstellungskraft, bin fantasievoll. <p>Antworten</p> <ul style="list-style-type: none"> • trifft überhaupt nicht zu • trifft eher nicht zu • weder noch • eher zutreffend • trifft voll und ganz zu <p>Fehlermeldung [WENN in mindestens einer Zeile keine Angabe gemacht wurde] Sie haben in mindestens einer Zeile noch keine Antwort gegeben. Bitte suchen Sie die entsprechende(n) Antwort(en) aus. Falls Sie keine Angabe machen möchten, klicken Sie bitte auf *Weiter*.</p>	<p>Question text Next, there are some general statements that can be used to describe a person. These statements can fit you more or less. How well do the following statements describe your personality? I see myself as someone who ...</p> <ul style="list-style-type: none"> • ... is generally trusting • ... has few artistic interests. • ... tends to find fault with others. • ... has an active imagination. <p>Answer options</p> <ul style="list-style-type: none"> • Disagree strongly • Disagree a little • Neither agree nor disagree • Agree a little • Agree strongly <p>Error message [IF no answer was given in at least one row:] You did not yet give an answer. Please select the respective answer. If you do not want to give an answer click on *Next*.</p>
Need for cognition	<p>Fragetext Die folgenden Aussagen können mehr oder weniger auf Sie zutreffen. Bitte geben Sie bei jeder Aussage an, inwieweit diese im Allgemeinen auf Sie persönlich zutrifft.</p> <ul style="list-style-type: none"> • Es genügt mir einfach die Antwort zu kennen, ohne die Gründe für die Antwort eines Problems zu verstehen. 	<p>Question text These statements can fit you more or less. Please choose for every answer how much it fits you personally.</p> <p>Answer options</p> <ul style="list-style-type: none"> • Simply knowing the answer rather than understanding the reasons for the answer to a problem is fine with me.

	<ul style="list-style-type: none"> • Ich habe es gern, wenn mein Leben voller kniffliger Aufgaben ist, die ich lösen muss. • Ich würde kompliziertere Probleme einfachen Problemen vorziehen. • In erster Linie denke ich, weil ich muss. <p>Antworten</p> <ul style="list-style-type: none"> • trifft überhaupt nicht zu • • • weder noch • • • trifft voll und ganz zu <p>Fehlermeldung [WENN in mindestens einer Zeile keine Angabe gemacht wurde] Sie haben in mindestens einer Zeile noch keine Antwort gegeben. Bitte suchen Sie die entsprechende(n) Antwort(en) aus. Falls Sie keine Angabe machen möchten, klicken Sie bitte auf *Weiter*.</p>	<ul style="list-style-type: none"> • I prefer my life to be filled with puzzles that I solve. • I would prefer complex to simple problems. • I think primarily because I have to. <p>Answer options</p> <ul style="list-style-type: none"> • Doesn't apply at all • • • Neither applies nor does not apply • • • Applies completely <p>Error message [IF no answer was given in at least one row:] You did not yet give any answer. Please select the respective answer. If you do not want to give an answer click on *Next*.</p>
Home-owner	<p>Frage Bewohnen Sie Ihre Wohnung/ Ihr Haus als ... ?</p> <p>Hilfetext Bei mietfreier Bereitstellung Ihrer Wohnung, zum Beispiel durch Familienmitglieder oder durch Ihren Arbeitgeber (Werk-/Dienstwohnung), sind Sie Hauptmieter/-in.</p> <p>Antworten</p> <ul style="list-style-type: none"> • Eigentümer/-in des Gebäudes • Eigentümer/-in der Wohnung • Hauptmieter/-in • Untermieter/-in <p>Fehlermeldung [WENN gar keine Angabe gemacht wurde:] Sie haben noch keine Antwort gegeben. Bitte suchen Sie die entsprechende Antwort aus. Falls Sie keine Angabe machen möchten, klicken Sie bitte auf *Weiter*.</p>	<p>Question text Do you live in your apartment as the ... ?</p> <p>Help text In case of rent-free apartment provision, for example by family members or your employer you are considered the main tenant.</p> <p>Answer options</p> <ul style="list-style-type: none"> • Owner of the building • Owner of the apartment • Main tenant • Sub-tenant <p>Error message [IF no answer was given:] You did not yet give an answer. Please select the respective answer. If you do not want to give an answer click on *Next*.</p>

Appendix C: Midpoint design experiment by panel

Table C1: Respondents in the middle category by experimental group and proportion tests based on z-scores by panel (absolute number of respondents by experimental group and p-values in parentheses)

	GIP	GP	Panel 1	Panel 2	Panel 3	Panel 4	Panel 5	Panel 6	Panel 7	Panel 8
5-point-scale										
% Conceptual&Visual midpoint	26.4	30.6	17.4	31.8	29.1	29.3	32.8	25.1	22.3	28.9
% Conceptual midpoint	26.4	29.7	17.8	27.3	24.7	27.2	26.8	32.1	23.4	31.8
% Difference	0.0	0.9	-0.4	4.5	4.4	2.1	6.0	-7.0	-1.2	-2.9
Z	0.0	0.4	-0.1	1.1	1.1	0.5	1.5†	-1.7	-0.3	-0.7
% Visual midpoint	8.3	11.3	6.3	13.5	10.9	14.7	15.0	19.0	13.8	16.1
% No midpoint	11.0	14.3	8.7	12.9	10.6	17.0	14.7	16.2	10.1	15.8
% Difference	-2.8	-3.0	-2.3	0.6	0.3	-2.2	0.2	2.8	3.6	0.2
Z	-1.8	-1.6	-1.0	0.2	0.1	-0.7	0.1	0.8	1.2	0.1
7-point-scale										
% Conceptual&Visual midpoint	15.1	14.4	17.5	17.2	17.5	18.0	17.0	21.0	16.9	12.2
% Conceptual midpoint	19.9	15.7	20.4	16	18.9	16.1	19.5	16.1	11.6	16.1
% Difference	-4.8	-1.3	-2.9	1.2	-1.3	1.9	-2.5	4.9	5.4	-3.9
Z	-2.5	-0.6	-0.8	0.4	-0.4	0.6	-0.7	1.4†	1.7*	-1.3
% Visual midpoint	13.0	11.1	9.4	13.2	7.9	7.3	12.0	12.9	8.9	10.8
% No midpoint	11.4	9.3	11.6	10.8	12.7	10.9	9.8	13.9	11.1	11.8
% Difference	1.6	1.8	-2.1	2.4	-4.8	-3.6	2.2	-1.0	-2.1	-1.0
Z	0.9	1.1	-0.8	0.8	-1.8	-1.4	0.8	-0.3	-0.8	-0.4

*** p<0.001; ** p<0.01; * p<0.05; †<0.1

General conclusion

General conclusion

Survey data are relevant because they can shape societal debates and can have an impact on policy decisions. However, for survey results to be informative it is necessary to take into account and ensure representativeness and response quality. In the following, I briefly summarize the findings from each dissertation paper and the conclusions that can be drawn from them. In addition, I depict paths for future research on the basis of this dissertation and I conclude with a summary of my overall research contribution.

Paper 1: Is there an association between survey characteristics and representativeness? A meta-analysis

In the first paper, I synthesize the existing literature on measuring survey representativeness and I assess common associations between survey characteristics (such as the sampling type and mode of data collection) and representativeness (as measured using R-Indicators and benchmark comparisons). I find that probability-based samples, mixed-mode surveys, and other-than-Web mode surveys are more representative than nonprobability samples, single-mode surveys, and web surveys. In addition, I find that there is a positive association between representativeness and the response rate. I conclude that there is an association between survey characteristics and representativeness and these results are partly robust across two common representativeness measures. There is, however, a strong need for more primary research into the representativeness of different types of surveys.

Future primary research is necessary to be able to enhance meta-analytic assessments of associations between survey characteristics and representativeness. More research into assessing survey representativeness on more than one measure would be especially valuable, as it would allow comparisons of results across measures. When more primary research is available, more advanced modeling can be applied in future meta-analyses. Multivariate analyses might, for example, provide a deeper understanding of the associations between survey characteristics and representativeness. In

addition, more primary research would enable exploring associations of survey representativeness with other potentially relevant survey characteristics (such as whether post-survey adjustments were applied or not).

Paper 2: Is it in the method? Testing five measures of survey representativeness

In the second paper, I compare five common measures of survey representativeness and assess their informative value in the context of representativeness comparisons within as well as across two probability-based online panels in Germany: the GIP and the GESIS Panel. Both panels share a number of survey design characteristics but also differ in other aspects. I assess the informative value of each representativeness measure in our study (response rates, R-Indicators, Fractions of Missing Information, subgroup response rates, and benchmark comparisons) for comparisons across and within the two probability-based online panels in Germany. I find that all five measures have advantages and disadvantages and they all shed light on different aspects of survey representativeness. Therefore, the extent to which these representativeness measures lend themselves to comparative analyses depends on the purpose of the investigation. I conclude from this study that for survey comparative representativeness analyses it is advisable to apply at least one measure at the aggregate level (response rates or, preferably, R-Indicators) in addition to at least one measure at the variable or category level (for example, subgroup response rates or benchmark comparisons) to obtain a comprehensive picture.

There are many possibilities for future research into the robustness of these findings in other settings. For instance, this paper focuses on a comparison of two panels that have the same target population. Whether the findings translate into the contexts of countries other than Germany and how valuable the examined measures of representativeness are in cross-national comparisons should, for example, be explored in future research. In addition, explorations into how the results on the measures examined in this paper change across panel waves might provide further valuable insights for researchers interested in assessing representativeness in longitudinal research. It might,

for instance, be that some measures disclose increases or decreases in representativeness across panel waves that other measures do not capture. Furthermore, there is generally little research into comparing different measures of representativeness, especially in observational studies. Therefore, to extend this study to other representativeness measures might be another relevant path for future research.

Paper 3: The utility of auxiliary data for survey response modeling: Evidence from the German Internet Panel

In the third paper, I take a closer look at the value of commonly used types of auxiliary data. I examine the utility of different sources of auxiliary data (sampling frame data, interviewer observations, and micro-geographic area data) for modeling survey response in a probability-based online panel in Germany. I explore whether auxiliary data are systematically missing by survey response. In addition, I investigate the correlations of the auxiliary data with survey response as well as the predictive power and the significance of coefficients of the auxiliary data in survey response models. I find that all of these data have disadvantages (for example scarcity, missing values, transparency problems, or high levels of aggregation) and none of them predict survey response to any substantial degree. I conclude that more research into the quality and predictive power of similar and other types of auxiliary data is needed to allow meaningful application of auxiliary data in survey practice, as for example in measuring representativeness, monitoring fieldwork, nonresponse adjustment, or conducting responsive design surveys.

Future research should, however, not only concentrate on searching for more auxiliary data that are predictive of survey response, but in addition examine more closely the problems with the already available auxiliary data. A relevant question for future research would, for instance, be why there is a substantial amount of missing data on the interviewer observations and why there is systematically more missing data for the nonrespondents than the respondents. This research could potentially be used to inform survey agencies, for example regarding interviewer trainings and monitoring. Another

question for future research would be whether some commercial and official micro-geographic area data lack predictive power regarding survey response models because of the high levels of aggregation. If so, it might be worthwhile to explore ways in which the error introduced into the data due to aggregation can be reduced.

Paper 4: Response quality in nonprobability and probability-based online panels

In the last paper of this dissertation, I shift the methodological focus to the measurement part of the TSE framework. In this paper, I investigate response quality in a comprehensive study of seven nonprobability online panels and three probability-based online panels, which were all collected in Germany during the same fieldwork period. In the analysis, I apply three response quality indicators: straight-lining in grid questions, item nonresponse, and midpoint selection in a visual design experiment. I find that there is significantly more straight-lining in the nonprobability online panels than in the probability-based online panels with regard to straight-lining. However, I find no systematic pattern indicating that response quality is lower in nonprobability online panels than in probability-based online panels with regard to item nonresponse and midpoint selection. I conclude that there is a difference between nonprobability online panels and probability-based online panels in response quality on one out of three satisficing indicators. Therefore, more research into response quality in nonprobability online panels is needed, for example using different indicators of response quality and different theoretical frameworks.

Future research might, for instance, investigate whether the significantly higher proportion of straight-lining can be replicated in other studies. In addition, it might be worthwhile to assess whether there are differences in other response quality indicators, such as response times, across nonprobability online panels. Furthermore, research into the association between sample representativeness and response quality would be valuable to assess data quality in nonprobability online panels and probability-based online panels from a broader TSE perspective.

Overall, with this dissertation I contribute to survey methodological research into the representativeness and response quality of survey data in several ways. I synthesize the existing literature on survey representativeness and I identify survey characteristics that are associated with survey representativeness. In addition, I demonstrate that survey representativeness needs to be examined from different perspectives using representativeness measures on different levels of aggregation. I also discover a lack in quality and predictive power of auxiliary data for survey response models and I explore differences in response quality across nonprobability and probability-based online panels. Furthermore, I identify a number of gaps in the current literature on representativeness and response quality of survey data in general.

The findings from this dissertation lead to the conclusion that great care has to be put into measuring and ensuring high survey data quality and more research is needed to fully understand how high representativeness and response quality can be reached. To undertake this research is imperative, because if survey data quality is compromised, research findings based on the data can be misleading. In future research, I therefore plan to expand on the work in this dissertation and to contribute to closing the identified gaps in the literature.

General Appendix: Eidesstattliche Erklärung

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

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

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