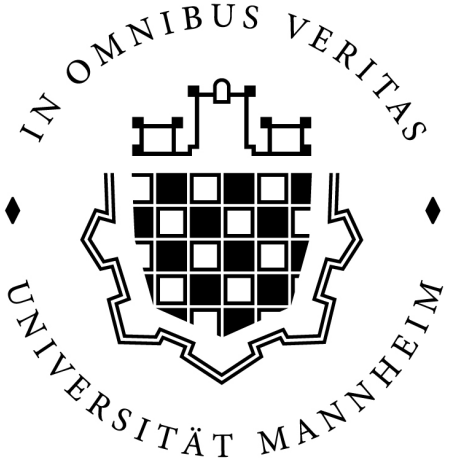
The impact of technological change on survey nonresponse and measurement

Jessica M. E. Herzing



The impact of technological change on survey nonresponse and measurement

Jessica M. E. Herzing

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

vorgelegt von
Jessica M. E. Herzing

Mannheim, 30.01.2018

Dekan der Fakultät für Sozialwissenschaften:
Prof. Dr. Michael Diehl

Erstbetreuerin:
Prof. Dr. Annelies G. Blom

Zweitbetreuerin & Erstgutachterin:
Prof. Dr. Frauke Kreuter

Zweitgutachterin:
Prof. Dr. Edith D. de Leeuw

Drittgutachter:
Prof. Dr. Florian Keusch

Tag der Verteidigung:
28.05.2018

List of papers

I Modeling group-specific interviewer effects on nonresponse using separate coding for random slopes in multilevel models

II The influence of a person's IT literacy on unit nonresponse and attrition in an online panel

III Investigating alternative interface designs for long-list questions

IV How do respondents use combo boxes? An evaluation

Acknowledgments

During the years I worked on this dissertation I received support in various ways from many people. I express my gratitude to all of them. First and foremost I would like to thank my advisors Annelies Blom and Frauke Kreuter for their invaluable personal and academic support, the time they spent in discussing my examination, and their patience. I would also like to thank the members of my advisory committee Edith de Leeuw and Florian Keusch for giving of their time to review and comment on my work. Furthermore, I would like to express my gratitude to Silke Schneider for encouraging me when my ideas for this dissertation were still emerging.

I am indebted to my current colleagues at the German Internet Panel and former colleagues at GESIS for constructive discussions, advice, and support; notably Barbara Felderer, Carina Cornesse, Christian Bruch, Daniela Ackermann-Piek, Franziska Gebhard, Julian Axenfeld, Klaus Pforr, Marina Fikel, Marina Jesse, Susanne Helmschrott, Tobias Rettig, Roberto Briceno-Rosas, Ulrich Krieger, and Verena Ortmanns.

The research presented here was supported by a number of grants and people to whom I owe an enormous amount of gratitude. I am beholden to Silke Schneider for allowing me space in the CAMCES pretest and supporting the proposal for the SOEP-IS experiment. I have to thank the SOEP-IS team for allowing me the implementation of my split-ballot experiment in their survey.

I have neglected my family and friends for many months, thank you for your patience, support, and care. A large part of the successful completion of this dissertation I owe to my parents and grandparents, who encouraged my educational aspirations. I am especially grateful to Steffen, who constantly reminded me that there is a world beyond this dissertation, who always had words of encouragement and who believed things would work out for the best (even when I was sure they would not).

Last but not least, I would like to thank the many respondents of the GIP, the SOEP-IS and the GESIS Pretest Lab: without their willingness to participate in the surveys and pretests this research would not have been possible.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Conducting surveys in a changing world

In recent decades, survey research has been strongly influenced by the digital change and human adaptation to technology. That survey research is affected by both the digital change and human adaptation to technology was already suggested by Dillman et al. (1998, p.1) in 1998, who stated [that] [c]omputer literacy varies greatly among people, as does the processing power of their computers. Thus, humans adapt differently to technology which results in differences in IT literacy (Antoun, 2015; van Deursen and van Dijk, 2014). Furthermore, human adaptation to changes in technology is slower than the change in the technology itself (Dillman, 2017). For example, humans differ in their device usage, and the devices themselves are from different technological generations. Accordingly, survey researchers have to make sure that when they adopt the survey design to technological changes that they do so without being too far ahead nor behind their target population.

On the one hand, the digital change has an impact on data collection modes. For example, in 1997 the first online surveys became feasible and practical, due to the increase of internet penetrations rates, increasing computer processing speed, and falling costs of hardware and software (see Hilbert and López, 2011). With the turn of the 21st century, there was a transition from collecting data with desktop computers to collecting data with laptops. Other forms of data collection emerged with the release of the first mass-market smartphones and tablet computers in 2007 and 2010 (Arthur, 2012; Sarwar and Soomro, 2013). Currently, the volume of data, the variety of data and the velocity with which we produce data, provide new opportunities for survey research to collect data, process data, and optimize measurements (e.g., GPS, movement, light, sound; for more information see Foster et al., 2017). Thus, survey research is adapting to technological advancements with regard to collecting data and optimizing measurements. For example, by tracking paradata, applying adaptive designs for different devices, or applying big data analytics. Consequently, the digital change is in full swing in the field of survey research.

On the other hand, survey research is confronted with another powerful develop-

ment namely human adaptation to technology. Internet penetration rates are increasing in Western-Europe and North America (Broadband Commission for Substainable Development, 2017). Internet penetration rates vary between 80.2 percent in Europe and 88.1 percent in North America in 2017 (Broadband Commission for Substainable Development, 2017). However, there is still a minority of people who do not have internet access (see European Commission, 2014). Furthermore, people differ in their access to the internet, such as having mobile internet access and/or having home internet accesses (European Commission, 2014, p. 44). For example, in 2014 46 percent of German households had a mobile and home internet access, 30 percent of German households had only home internet access, 3 percent had only mobile internet access, whereas 21 percent had no internet access at all (European Commission, 2014, p. 44). These differences in access to the internet might influence coverage error and sampling error (for examples see Blom et al., 2017; Bosnjak et al., 2013; de Vos, 2010; Knoef and de Vos, 2009; Revilla et al., 2016). For example, younger and better-educated people are more likely to use the internet than elderly and low educated people. Hence, the exclusion of people without internet might result in coverage biases (Blom et al., 2017).

Moreover, respondents differ in their usage of the internet and the devices they use (Antoun, 2015; van Deursen and van Dijk, 2014). These differences in the use of devices is of particular interest to survey research because it is likely associated with nonresponse to online surveys (for the association between response rates and mode preferences see Millar and Dillman, 2011; Olson et al., 2012; Rookey et al., 2008; Shih and Fan, 2007). For example, respondents who only use the internet via smartphone might be unwilling to use a computer to fill out an online questionnaire. In addition, the usage of different devices, such as smartphones or PCs, has an impact on measurement (Antoun et al., 2017). In this regard, Lugtig and Toepoel (2016) showed that there is a higher measurement error of respondents using tablets and smartphones. Survey practitioners have to consider this heterogeneity among sample units' technology usage, as differences in technology usage might affect various types of survey errors (coverage, sampling, nonresponse, measurement error) and hence, overall data quality.

Likewise, it is important that survey methodologists gain a deeper understanding of how technological advancements affect survey errors. With this understanding survey researchers can then optimize their data collection and produce high-quality data. This dissertation addresses the challenges to computer-assisted self-administered interviews

posed by recent technological changes. For this purpose, I investigate the influence of respondents' technological adaptation on online survey participation, and I evaluate alternative interface designs for complex survey questions in computer-assisted surveys.

This dissertation uses the survey lifecycle as its framework. The survey lifecycle allows to understand each step required for a successful data collection and preparation, and helps to identify errors associated with different steps. Within the framework of the survey lifecycle, special emphasis will be put on nonresponse and measurement. The four studies that constitute this dissertation cover the design perspective and the quality perspective of the survey lifecycle. Accordingly, this dissertation aims at advancing the understanding of the causes of errors and the influence of survey design decisions in order to maximize data quality.

1.2 The survey lifecycle

Researchers need high quality data to make valid statistical inference. To increase data quality survey researchers strive for low variance and low bias. To achieve this goal survey researchers try to understand why errors (variance and bias) arise in survey statistics. Errors can occur when survey researchers do not measure what they intended to measure (the measurement side of the survey lifecycle) or the people who participate in the survey are not representative of the target population (the representation side of the survey lifecycle). One way of understanding the sources of errors is the study of survey design decisions that are made when designing a survey (Fowler, 1993).

To study potential pitfalls in surveys, researchers investigate why errors occur which is called "design perspective" (see Groves et al., 2011, p. 41). To examine and quantify specific errors is characteristic of the "quality perspective" (see Groves et al., 2011, p. 41). Groves et al. (2011, ch. 2) combined these two perspectives into one framework – the survey lifecycle – to measure and minimize errors in survey statistics (for an illustration see figure 1.1).

The *design perspective* describes a set of decisions that survey practitioners need to make when designing surveys (represented by the boxes in figure 1.1). The left side of figure 1.1 presents aspects which are related to the measurement. When survey researchers design measurements to gather information from respondents (measurement

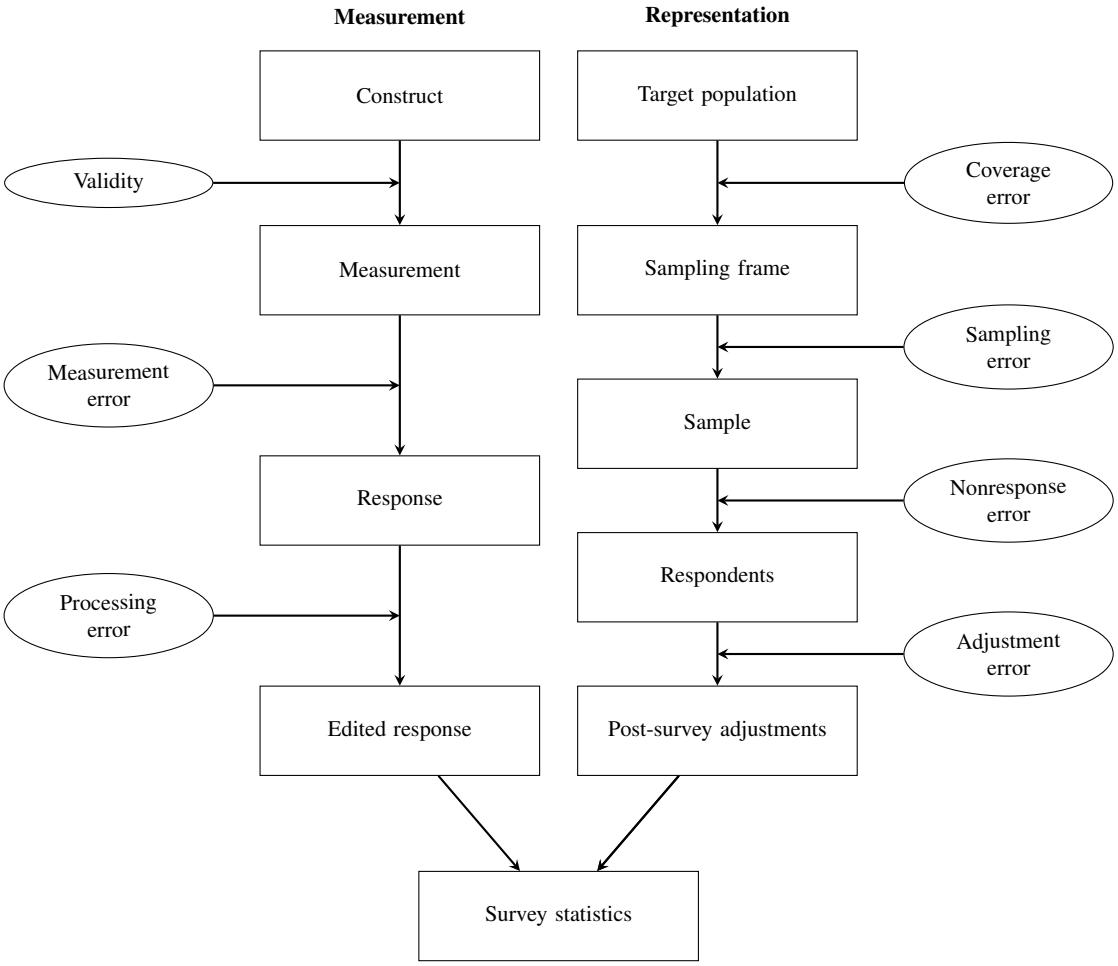


Figure 1.1: Survey lifecycle based on Groves et al. (2011, p. 48).

side of the survey lifecycle) they first define the construct they intend to measure, e.g., measuring education. At this measurement design stage, the wording of the question is relatively abstract and does not exactly describe what is meant by the construct (Groves et al., 2011, p. 42). After that, survey researchers design a question which measures the construct and which results in a measurement. These measurements are used to collect responses and hence, measurements gather information from respondents, that produce data. Finally, the response is edited by respondents themselves. Alternatively, researchers can also edit the response in computer-assisted surveys by implementing checks in terms of the entered response format, e.g., whether respondents gave a number when asked for the year of birth. All survey design decisions that are made within these steps of the measurement side can affect the survey statistics.

The right side of figure 1.1 describes the representation side of the survey lifecycle. Initially, survey researchers decide which people they intend to study by defining the target population. Afterward, the units of the target population are listed within the sampling frame. Subsequently, a sample is selected from the sampling frame. Then, researchers attempt to take measurements from the selected sample. The sample units which were successfully measured are called respondents (or nonrespondents or unit nonrespondents). Finally, researchers can develop post-survey adjustments to improve the quality of the estimates by examining the unit nonresponse of different respondent groups. All survey design decision on the representation side of the survey lifecycle can affect survey statistics.

The *quality perspective* of the survey lifecycle covers the Total Survey Error (TSE) paradigm (represented by ovals in figure 1.1). The TSE is an umbrella term covering multiple types of errors that may occur when conducting population surveys. There are six survey error components covered by the TSE paradigm which sum up to the total survey error (see Biemer and Lyberg, 2003; Groves et al., 2011, p. 48). Each of these quality components has a statistical formulation and is a property of individual survey statistics (Groves et al., 2011, p. 49).

The measurement side begins with validity which is the difference between the construct and the measures. Next, we can estimate the measurement error which is the difference between an ideal measurement and the actual response. The measurement error consists of response variance (or reliability) and response bias. Finally, the measurement side covers the processing error which is the difference between the response

given by the respondents and the value used in the estimation.

The representation side starts with coverage error which is the difference between the target population that is covered and that is not covered by the sampling frame. If those who are not covered by the sampling differ from the target population, then coverage bias can occur. The sampling error is introduced when not all sample units of the sampling frame are measured because of costs or infeasibility. The sampling error consists of sampling bias and sampling variance. Sampling bias emerges when there is a systematic failure to observe all people from the sampling frame, and when these people have different characteristics. However, even when there is no systematic failure of being selected the same sample design can result in different samples which is called sampling variance (Groves et al., 2011, p. 98). When not all sample units are successfully measured (nonresponse variance), and the respondents' data differs from the entire sample researchers speak of nonresponse bias. The final point in the TSE paradigm is the adjustment error which occurs when researchers adjust for nonobservations. The statistical quality of the survey estimates is influenced by each of the seven error components.

The current state of research on how to make design and estimation decisions when technological innovations are implemented in the survey design to minimize error in survey statistics is evolving but far from complete. The primary purpose of my research is the investigation of whether respondents are willing and able to respond to the survey request and survey questions in a technologically changing survey world from a survey lifecycle perspective.

I cover the quality perspective of the survey lifecycle with two papers on reasons for nonresponse errors in a probability-based online panel (chapter 2 and chapter 3). Two studies of this dissertation place emphasis on the survey design perspective. For this purpose, the third (chapter 4) and the fourth paper (chapter 5) evaluate the measurement quality of two alternative interface designs for long-list questions.

1.2.1 Nonresponse error

Out of the seven error components of the TSE paradigm, I will only discuss the nonresponse error. More specifically I only cover the aspect of nonresponse bias. In general, nonresponse is survey researchers failure to get information from sample units (Groves

et al., 2011, p. 183). When survey researchers completely miss information from sample units, they speak about unit nonresponse. Nonresponse rates are the percentage of eligible sample units that have not been measured (expressed as $\frac{N-n}{N}$ where N is the full sample and n are the respondents). When there is a systematic difference in characteristics y between respondents and nonrespondents, then nonresponse bias can occur. The deterministic view of nonresponse bias defines the mean of nonresponse bias as

$$\text{Nonresponse bias}(\bar{y}) = (\frac{N-n}{N}) * (\bar{y}_r - \bar{y}_{nr}), \tag{1.1}$$

where $\frac{N-n}{N}$ is the nonresponse rate, $\bar{y}_r$ is the mean of respondents, and $\bar{y}_{nr}$ is the mean of nonrespondents. This expression exemplifies that the higher the nonresponse rate, the higher the nonresponse bias. This model was transformed into a stochastic view by Bethlehem (2002, p. 276) that assumes that each sample unit is a potential respondent or a potential nonrespondent. Accordingly, Bethlehem (2002, p. 276) considers in his expression the response propensities (the probability of participation of sample units, ρ) which changes expression 1.1 into

$$\text{Nonresponse bias}(\bar{y}) = \frac{Cov_{y,\rho}}{\bar{\rho}}, \tag{1.2}$$

where $Cov_{y,\rho}$ is the covariance between the variable of interest (y) and the propensity to respond (ρ) among the sample units; and $\bar{\rho}$ is the mean propensity to respond across all sample units (equivalent to response rate). The covariance is a function of the correlation of y and ρ multiplied by the variance of y multiplied by the variance of ρ. If one of these three components of the covariance is zero, then there is no nonresponse bias.

Expression 1.1 shows that nonresponse bias is only partly a function of nonresponse rates (see Brick and Tourangeau, 2017). Furthermore, expression 1.2 indicates that nonresponse bias can vary within one survey because bias depends on whether the response propensity and the survey variables of interest are correlated (Groves and Peytcheva, 2008).

Because response rates might affect nonresponse bias, we need to understand why people do not want to participate in surveys anymore (Tourangeau, 2017). In this context, Groves et al. (2011, p. 202) suggest several aspects for reducing unit nonresponse at sample units' initial decision to participate in a survey: interviewer behavior, sponsorship, pre-notification, incentives, response burden, respondent rule, household/inter-

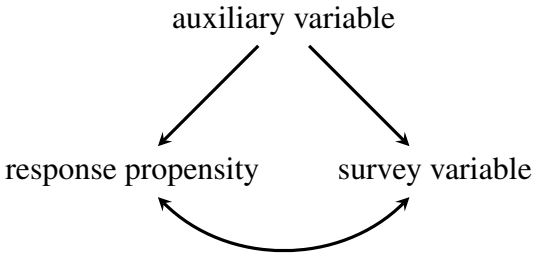


Figure 1.2: Common cause model based on Groves (2006).

viewer match. In the first paper of this dissertation we investigate interviewer effects because interviewers have an impact on the sampling frame, coverage, contact, and recruitment of potential respondents, survey measurements, and data processing (for more information see West and Blom, 2016). More specifically, interviewers often play a central role in potential nonresponse bias (Loosveldt and Beullens, 2014; West and Olson, 2010). In the first paper of this dissertation, we focus on interviewers' impact on unit nonresponse to a probability-based online panel (see chapter 2). More specifically, we investigate whether the size of interviewer effects on nonresponse differs across onliners (internet households) and offliners (non-internet households).

The leading article on the relationship between response propensity and nonresponse bias by Groves (2006) suggests three alternative causal models: separate cause model, survey variable cause model, and common cause model. The separate cause model states that the causes of a variable are independent of the causes of response propensity (Groves et al., 2011, p. 191). The survey variable cause model says that a survey variable itself is a cause of response propensity. According to the common cause model (illustrated in figure 1.2), nonresponse bias can occur if there are auxiliary variables that affect both persons' response propensity as well as the survey variables of interest to the researcher (shared causes); and if response propensity and survey variables are also associated with each other. The corresponding question is which underlying factors cause the associations between response propensity and survey variables. In the second paper, we build on the common cause model by investigating whether IT literacy is an underlying factor for unit nonresponse in a probability-based online panel.

1.2.2 Measurement

Measurements are ways to gather information from respondents. The measurements most often used in surveys are questions which can be communicated orally or visually. There are several theories that describe the process of respondents answering a survey question (e.g., Petty and Jarvis, 1996). For this purpose, I build on the theoretical foundation of the conceptualization of the question-answering process by Tourangeau et al. (2000) who specified four main components of the response process: comprehension, retrieval, judgment, and response. This question-answering process was extended with the component "perception" for self-administered surveys by Jenkins and Dillman (1997). Understanding the question-answering process helps survey researchers to identify sources of errors in the measurement.

When confronted with a survey question, respondents *perceive* the question, which means they see the question which has been asked. In this context, the respondent decides which part of the page/screen he/she wants to focus on (Jenkins and Dillman, 1997). When picking and interpreting the relevant stimuli, the respondents use their previous knowledge. Out of this knowledge, they build a pattern on which parts of the page/screen they pay attention to. Thus, there is a visual communication between researchers and respondents.

Next, respondents try to *comprehend* what is meant by the question. A critical aspect of this stage is that respondents understand the meaning of each word and the question as a whole in the same way as researchers intended. Respondents try to interpret researchers' intentions, in doing so they use verbal and nonverbal cues (e.g., words, spacing, scrollbars). Thus, respondents need to comprehend the visual cues as well as the wording as it was intended by the researcher (see Jenkins and Dillman, 1997; Tourangeau et al., 2004).

During the *retrieval* process, respondents search their memories for information necessary to answer the question. Depending on the question either factual memories (in case of fact questions) or judgmental memories (in case of attitude questions) are awakened (see Biemer and Lyberg, 2003; Schwarz and Sudman, 1996).

After retrieving information, a *judgment* is made to generate a response. In this step, respondents assess whether their memories are complete and relevant to answer the survey question.

Finally, respondents edit their *response* to the answer format presented. The respondents' answer has to be adjusted or formatted with regard to the answer options given. Nonverbal and verbal cues influence this answer formation process independently or jointly (Tourangeau et al., 2000, p. 13). For example, Tourangeau et al. (2004) showed that the placement of nonsubstantive response options (e.g., "Don't know" category) in relation to the substantive options affects the interpretation of the conceptual and the visual midpoints of scales.

In self-administered questionnaires, the five components of the question-answering process are influenced by the interface design of the question (Schwarz, 1999). The question-answering process affects the accuracy of responses (for the example of autobiographical questions see Bradburn et al., 1987). Furthermore, respondents differ in their cognitive efforts when answering the questions, and thus the question-answering process differs between respondents and sometimes within respondents (see Schwarz and Sudman, 1996; Tourangeau et al., 2000). For example, respondents may select a response that is adequate rather than optimal (for more information on response order effects see Krosnick and Alwin, 1987; Tourangeau et al., 2000, p. 250 ff.). It is the researcher's aim that respondents choose the optimal rather than adequate response option when answering a question. Accordingly, gaining knowledge of the question-answering process is important for evaluating whether an interface design makes the task of reporting an answer easier for respondents or not. When it is easy for respondents to answer the survey question, then the chance of getting exact measures from respondents increases. This association is driven by the assumption that questions which need less cognitive effort to answer introduce less measurement error than questions that need more cognitive effort to answer (Tourangeau et al., 2000).

To investigate whether the measurement allows an optimal question-answering process survey researchers test and evaluate survey questions in pretests. When conducting pretests researchers test and evaluate whether problems emerge in the components of the question-answering process (Tourangeau et al., 2000). To optimize the question-answering process Groves et al. (2011, p. 259) proposed three standards that the measurement should meet:

- Content standards: Respondents understand the questions as intended by the research.

- Cognitive standards: Respondents can understand and answer the question with the information given.
- Usability standards: Respondents can use the survey instrument and can report their answer in the provided response format.

One option to evaluate survey questions based on the aforementioned three standards is the experimental comparison of different versions of questions or interface designs with split-ballot experiments (Presser et al., 2004; Tourangeau, 2004). In the third paper, I use a split-ballot experiment to evaluate whether different question instruments for long-list questions produce different answers and whether different question instruments for long-list questions result in different response burdens (see chapter 4). Split-ballot experiments offer the potential to evaluate the impact of proposed design changes on the question-answering process and the resulting data which other pretesting approaches do not provide (Groves et al., 2011, p. 268).

In the fourth paper, we use cognitive interviews in combination with an eye-tracking study to evaluate an alternative interface design for long-list questions, namely combo boxes (a combination of a text field and a drop-down box, see chapter 5). With this study, we identify usage problems of combo boxes which cause difficulties in the question-answering process of respondents. In this regard, we investigate cognitive standards and usability standards.

1.3 Summaries of the papers

This dissertation examines survey design choices with regard to measurement and sources of errors with regard to unit nonresponse in times of rapid digital change. The four papers in this dissertation deal with respondents' ability to use technology and how respondents' technological abilities are associated with nonresponse and measurement. In the first part of this dissertation, I investigate the association between respondents' technology adaptation on nonresponse to a probability-based online panel (chapter 2 and chapter 3). The findings of paper I and II are useful for researchers who want to reduce nonresponse within the fieldwork process or who want to develop post-survey adjustments. In the second part of this dissertation, I evaluate alternative interface designs for

long-list questions in computer-assisted surveys (chapter 4 and chapter 5). The results of the third and fourth paper are relevant for survey practitioners who aim to maximize response quality for questions with long lists of response options. Optimized versions of the interface designs, which were developed and tested in this dissertation, are already implemented in a number of large surveys, such as the SOEP Migration Sample, the SOEP Refugee Sample and the ReGES (Refugees in the German Education System) for the measurement of educational qualification. In the following, I give a summary of each paper.

Paper I In paper I (chapter 2) we study the impact of interviewers on the recruitment of onliners (internet households) and offliners (non-internet households) in a probability-based online panel. The central question is whether the size of interviewer effects on nonresponse is different for onliners and offliners. To investigate this question, we estimated a multilevel model with random slopes to estimate interviewer effect size with data from the face-to-face recruitment interview of the German Internet Panel (GIP). We propose an alternative parametrization for the investigation of interviewer effect sizes namely separate coding in the random slopes for the groups of offliners and onliners. Our results show smaller interviewer effects in the recruitment of offliners than in the recruitment of onliners. The smaller interviewer effects indicate that the low response rates of offliners are not associated with interviewers' recruitment strategies. Nevertheless, interviewers are unequally good in recruiting onliners. While some interviewers are more successful in recruiting onliners, some interviewers are less successful in recruiting onliners to the online panel. Overall, we learn from the first paper that we cannot raise the response propensities of offliners by introducing interviewer-related fieldwork strategies (e.g., target specific respondents). The importance of the proposed parametrization strategy lies in its information on whether it is efficient to increase the effort in interviewer-related fieldwork strategies for hard cases or not.

Paper II Paper II (chapter 3) explores the diversity of subgroups among the internet population and why different subgroups of the online population are hesitant to participate in an online panel. We scrutinize whether persons' IT literacy is a predictor for nonresponse bias in the GIP. We find that persons belonging to different classes of IT literacy have systematically different socio-demographic characteristics and show dif-

ferent voting behavior. In addition, we find that initial response propensities vary by classes of IT literacy, as do retention patterns over time. Thus, not only offliners participate less in the online panel, but also specific subgroups of the online population. Understanding the mechanism leading to selectivities in the data of online surveys offers new opportunities of adjusting for such selectivities resulting in better population estimates which is a prerequisite for making substantive research count. In addition, this research indicates that even though the internet penetration rates increase in Western Europe and North America, there are differences in online survey participation with regard to respondents adaptation to technology.

Paper III Paper III (chapter 4) addresses the scope for interface design and its implications on measurement in computer-assisted surveys. I assessed, whether and how interface designs support respondents in answering questions. Furthermore, I evaluated whether alternative interface designs come along with a change in data quality. For this purpose, I developed three different response formats for the question on the highest educational qualification in Germany: a long list of radio buttons, a combo box (also called lookup database), and a search tree. To evaluate the interface designs, I implemented a split-ballot experiment in the Innovation Sample of the Socio-Economic-Panel (SOEP-IS, experiment was implemented after my application was successful in the referee process). To evaluate the measurement quality of the three interface designs, I investigated response burden and data quality. My results indicate a decrease in response burden for the combo box condition. However, the combo box condition results in more post-coding than search trees and long lists. My results suggest that there is a trade-off between a decrease of response burden and an increase of post-survey coding across the interface designs. Hence, survey practitioners have to decide on a question specific basis whether they want to implement a combo box or not.

Paper IV The last paper (chapter 5) tackles the issue of measurement quality with regard to cognitive and usability standards. For this purpose, we designed cognitive interviews in combination with an eye-tracking study to investigate how respondents use the interface design of a combo box (a combination of a text field and a drop-down box) and whether the usage of the combo box comes along with specific response difficulties. With the help of the eye movements, we detect that respondents use the combo box

slightly more often like a drop-down box than as a standard text field. We find that both usage types are associated with specific problems in the question-answering process. Nevertheless, the vast majority of respondents said that the combo box was easy to use and that they found their intended answer. Our usability testing has further implications for the optimization of the interface design of combo boxes and its corresponding search algorithm. For example, more visual cues (e.g., magnifying glass) are needed to draw respondents' attention to the search function of the combo box.

Collectively, my findings suggest that survey designs influence respondents on different aspects of the survey lifecycle, such as nonresponse and measurement. However, the impact of survey design varies by respondents adaptation to technology. Even though the adaptation to technology increases, some respondents are reluctant to participate in an online survey which might introduce nonresponse bias. In addition, I observe that measurements can be improved by technological advancements, such as long-list questions. However, respondents use alternative interface designs differently and not necessarily more efficiently. Subsequently, my dissertation demonstrates that survey researchers still have to be cautious when adapting the survey design to technological advancements because these technological advancements might be too far ahead of their respondents.

References

Antoun, C. (2015). Who are the Internet Users, Mobile Internet Users, and Mobile-Mostly Internet Users?: Demographic Differences across Internet-Use Subgroups in the U. S. In Toninelli, D., Pinter, R., and de Pedraza, P., editors, *Mobile Research Methods: Opportunities and Challenges of Mobile Research Methodologies*, chapter 7, pages 99–117. Ubiquity Press, London, UK.

Antoun, C., Couper, M. P., and Conrad, F. G. (2017). Effects of Mobile versus PC Web on Survey Response Quality: A Crossover Experiment in a Probability Web Panel. *Public Opinion Quarterly*, 81(S1):280–306.

Arthur, C. (2012). The History of Smartphones: Timeline. *The Guardian*. https://www.theguardian.com/technology/2012/jan/24/smartphones-timeline.

Bethlehem, J. G. (2002). Weighting nonresponse adjustments based on auxiliary information. In Groves, R. M., Dillman, D. A., Eltinge, J. L., and Little, R. J. A., editors, *Survey Nonresponse*, pages 221–251. Wiley, New York, NY.

Biemer, P. P. and Lyberg, L. E. (2003). *Introduction to Survey Quality*, volume 335 of *Wiley Series in Survey Methodology*. John Wiley & Sons, Hoboken, NJ.

Blom, A. G., Herzing, J. M. E., Cornesse, C., Sakshaug, J. W., Krieger, U., and Bossert, D. (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, M., Haas, I., Galesic, M., Kaczmirek, L., Bandilla, W., and Couper, M. P. (2013). Sample Composition Discrepancies in Different Stages of a Probability-Based Online Panel. *Field Methods*, 25(4):339–360.

Bradburn, N. M., Rips, L. J., and Shevell, S. K. (1987). Answering Autobiographical Questions: The Impact of Memory and Inference on Surveys. *Science*, 236(4798):157–161.

Brick, J. M. and Tourangeau, R. (2017). Responsive Survey Designs for Reducing Nonresponse Bias. *Journal of Official Statistics*, 33(3):735–752.

Broadband Commission for Substainable Development (2017). *The State of Broadband 2017: Broadband Catalyzing Substainable Development*. United Nations, ITU.

de Vos, K. (2010). Representativeness of the LISS-Panel 2008, 2009, 2010.

Dillman, D. A. (2017). The Worldwide Challenge of Pushing Respondents to the Web in Mixed-Mode Surveys. In *The 28th International Workshop on Household Survey Nonresponse*, Utrecht, NL.

Dillman, D. A., Tortora, R. D., and Bowker, D. (1998). Principles for Constructing Web Surveys. In *Joint Meetings of the American Statistical Association*, Baltimore, ML.

European Commission (2014). Special Eurobarometer 414, Wave EB81.1, E-Communications and Telecom Single Market Household Survey. http:

//ec.europa.eu/COMMFrontOffice/publicopinion/index.cfm/ResultDoc/download/DocumentKy/57810.

Foster, I., Ghani, R., Jarmin, R. S., Kreuter, F., and Lane, J., editors (2017). *Big Data and Social Science–A Practical Guide to Methods and Tools*. CRC Press, Taylor & Francis Group, Boca Raton, FL.

Fowler, F. J. (1993). *Survey Research Methods*. Applied Social Research Methods Series. Sage Publications, Los Angeles, London, New Dehli, Singapore, Washington DC, 4th edition.

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., and Tourangeau, R. (2011). *Survey Methodology*, volume 561 of *Wiley Series in Survey Methodology*. John Wiley & Sons, Hoboken, NJ, 2nd edition.

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.

Hilbert, M. and López, P. (2011). The World's Technological Capacity to Store, Communicate, and Compute Information. *Science*, 332(6025):60–65.

Jenkins, C. R. and Dillman, D. A. (1997). Towards a Theory of Self-Administered Questionnaire Design. In Lyberg, L. E., Collins, M., de Leeuw, E. D., Dippo, C., Schwarz, N., and Trewin, D., editors, *Survey Measurement and Process Quality*, Wiley Series in Probability and Statistics, chapter 7, pages 165–196. John Wiley & Sons, New York, NY.

Knoef, M. and de Vos, K. (2009). *The Representativeness of LISS, an Online Probability Panel*. CentERdata, Tilburg, NL. https://www.researchgate.net/profile/Marike_Knoef/publication/242742051_The_representativeness_of_LISS_an_online_probability_panel/links/0f3175339ae828f081000000.pdf.

Krosnick, J. A. and Alwin, D. F. (1987). An Evaluation of a Cognitive Theory of Response-Order Effects in Survey Measurement. *Public Opinion Quarterly*, 51(2):201–219.

Loosveldt, G. and Beullens, K. (2014). A Procedure to Assess Interviewer Effects on Nonresponse Bias. *SAGE Open*, 4(1):1–12.

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

Millar, M. M. and Dillman, D. A. (2011). Improving Response to Web and Mixed-Mode Surveys. *Public Opinion Quarterly*, 75(2):249–269.

Olson, K., Smyth, J. D., and Wood, H. M. (2012). Does Giving People Their Preferred Survey Mode Actually Increase Survey Participation Rates? An Experimental Examination. *Public Opinion Quarterly*, 76(4):611–635.

Petty, R. and Jarvis, W. (1996). An Individual Differences Perspective on Assessing Cognitive Processes. In Schwarz, N. and Sudman, S., editors, *Answering Questions*, pages 221–251. Jossey–Bass Publishers, San Francisco, CA.

Presser, S., Couper, M. P., Lessler, J. T., Martin, E., Martin, J., Rothgeb, J. M., and Singer, E. (2004). Methods for Testing and Evaluating Survey Questions. *Public Opinion Quarterly*, 68(1):109–130.

Revilla, M., Cornilleau, A., Cousteaux, A.-S., Legleye, S., and de 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.

Rookey, B. D., Hanway, S., and Dillman, D. A. (2008). Does a Probability-Based Household Panel Benefit from Assignment to Postal Response as an Alternative to Internet-Only? *Public Opinion Quarterly*, 72(5):962–984.

Sarwar, M. and Soomro, T. R. (2013). Impact of Smartphone's on Society. *European Journal of Scientific Research*, 98(2):216–226.

Schwarz, N. (1999). Self-Reports: How the Questions Shape the Answers. *American Psychologist*, 54(2):93–105.

Schwarz, N. and Sudman, S. (1996). *Answering Questions: Methodology for Determining Cognitive and Communicative Processes in Survey Research*. Wiley: Jossey-Bass, Hoboken, NJ.

Shih, T.-H. and Fan, X. (2007). Response Rates and Mode Preferences in Web-Mail Mixed-Mode Surveys: A Meta-Analysis. *International Journal of Internet Science*, 2(1):59–82.

Tourangeau, R. (2004). Experimental Design Considerations for Testing and Evaluating Questionnaires. In Presser, S., Rothgeb, J. M., Couper, M. P., Lessler, J. T., Martin, E., Martin, J., and Singer, E., editors, *Methods for Testing and Evaluating Survey Questionnaires*, Wiley Series in Survey Methodology, chapter 11, pages 209–224. John Wiley & Sons, New York, NY.

Tourangeau, R. (2017). Presidential Address Paradoxes of Nonresponse. *Public Opinion Quarterly*, 81(3):803–814.

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

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

van Deursen, A. J. A. M. and van Dijk, J. A. G. M. (2014). The Digital Divide Shifts to Differences in Usage. *New Media & Society*, 16(3):507–526.

West, B. T. and Blom, A. G. (2016). Explaining Interviewer Effects: A Research Synthesis. *Journal of Survey Statistics and Methodology*, 5(2):175–211.

West, B. T. and Olson, K. (2010). How Much of Interviewer Variance is Really Nonresponse Error Variance? *Public Opinion Quarterly*, 74(5):1004–1026.

Chapter 2

Paper I

Modeling group-specific interviewer effects on nonresponse using separate coding for random slopes in multilevel models

Abstract

While there is ample evidence of interviewers affecting nonresponse and some evidence regarding the factors explaining overall interviewer effects, the literature is sparse on how interviewers differentially affect specific groups of sample units despite the importance of this in terms of nonresponse bias. A reason for the sparse literature on interviewer effects on nonresponse bias may be limitations of standard use of multilevel models. We demonstrate how an alternative parametrization of the random components in multilevel models, so-called separate coding, delivers insights into differential interviewer differential effects on specific respondent groups. The multilevel model with separate coding of random coefficients makes it not only possible to estimate how the size of interviewer effects varies across types of respondents, but also offer possibilities to investigate how interviewer characteristics affect the groups differentially. Investigating nonresponse during the recruitment of a probability-based online panel separately for persons with and without prior internet access, we detect that the size of the interviewer effect differs between the two respondent groups. While we discover no interviewer effects on nonresponse for offliners, we find sizable interviewer effects for onliners. In addition, we identify interviewer characteristics that explain this group-specific nonresponse.

Modeling group-specific interviewer effects on nonresponse using separate coding for random slopes in multilevel models[1]

2.1 Introduction

A decline in response rates in population surveys over the past two decades (de Heer and de Leeuw, 2002) paired with potentially selective nonresponse (see McCabe and West, 2016) endanger unbiased inference from survey data. Surveys may have suspicions about key respondent groups that may be underrepresented depending on the survey's target population, mode, or topic salience. For example, probability-based online surveys seek unbiased responses in terms of households with and without internet access but are aware that offline respondent groups (non-internet households) may be less likely to respond to an online survey than onliners (internet households). Thus, to counter potential nonresponse biases, surveys need to monitor nonresponse among specific respondent groups and identify measures to enhance response among underrepresented groups.

To increase response rates and to decrease nonresponse bias survey practitioners often use interviewers in population surveys (Heerwegh, 2009). While interviewers tend to increase overall response rates in surveys (see Heerwegh, 2009), research on the determinants of nonresponse have also identified human interviewers as one reason for variations in response rates (Couper and Groves, 1992; Loosveldt and Beullens, 2014; West and Blom, 2016). In addition, we know from research on interviewer effects that interviewers introduce nonresponse bias, if they systematically differ in their success in obtaining a response from different respondent groups (West and Olson, 2010; West et al., 2013). Therefore, interviewers might be a source of selective nonresponse in surveys.

So far, research on interviewer effects on nonresponse has been primarily concerned with interviewers' differential response rates (see for example Durrant et al., 2010; Durrant and Steele, 2009; Hox and de Leeuw, 2002). However, interviewers might also

[1] This chapter is co-authored with Annelies G. Blom and Bart Meuleman.

differentially contribute to nonresponse bias, when the interviewer effect is correlated with the characteristics of the approached sample units. A notable exception to the general trend in the literature is Loosveldt and Beullens (2014), who investigate whether interviewer effects on nonresponse differ across specific respondent groups.

To investigate interviewer effects for specific groups of respondents, Loosveldt and Beullens (2014) estimated multilevel models with respondents nested within interviewers and included dummies in the random part of the model to distinguish between respondent groups. Dummy coding, which is also referred to as contrast coding (Jones, 2013), selects one category as the reference, and the dummy effects represent the contrast with this reference group (whose value is captured by the intercept). When such dummy variables are included as random components in multilevel models for interviewers effects, the obtained variance estimates indicate to what extent the contrast between respondent groups varies across interviewers. Yet, such parameterization does not directly yield insight into the relevant question of how large the interviewer effects are for respondent groups separately.

To obtain information on differences in interviewer effects (rather than interviewer variation in differences per interviewers), we propose to use an alternative coding strategy, so-called separate coding (Jones, 2013). In the case of separate coding, the intercept is omitted from the model, and as many 0/1 variables are included as there are categories. As a result, every variable represents a direct estimate for the group (rather than the contrast with a reference category).

When separate coding is used in the random part, one obtains a direct estimate of the interviewer effects for specific respondent groups. In multilevel models with random slopes the different coding strategies - contrast and separate coding - have an impact on the interpretation of the variance components and thus, answer different research questions.

In the following, we illustrate how researchers can analyze group-specific interviewer effects using different coding strategies for sample unit characteristics in multilevel models with random slopes - contrast coding versus separate coding. We investigate the face-to-face recruitment interview of the probability-based German Internet Panel (GIP). We know that the GIP data suffer from group-specific nonresponse, as response rates among onliners and offliners differ significantly (Blom et al., 2017). This paper demonstrates how separate coding in multilevel models with random slopes - as an

alternative to the more widely used contrast coding - can detect and explain differences in nonresponse among onliners and offliners.

In summary, this paper sets out to answer the following research questions:

(1) To what extent do interviewers affect nonresponse to a probability-based online panel?

(2) Is the size of interviewer effect on nonresponse different for onliners and offliners?

(3) If so, which interviewer characteristics differentially influence interviewer effects on nonresponse among onliners and offliners?

The importance of the interviewer effect size lies in its prediction of the effectiveness of different interviewer-related fieldwork strategies (for examples on liking, matching, or prioritizing respondents with interviewers see Durrant et al., 2010; Peytchev et al., 2010; Pickery and Loosveldt, 2004, 2002). Group-specific interviewer effect sizes indicate to which extent the implementation of interviewer-related fieldwork strategies is effective for the reduction of nonresponse in specific respondent groups. If the interviewer effect size is the same across respondent groups, optimal interviewer strategies for reducing nonresponse will have similar effects across all groups. However, if the difference in interviewer effect size is large between specific respondent groups, some interviewer strategies may be efficient for some groups of respondents but not for all. Therefore, understanding group-specific interviewer effect sizes can aide the efficiency of respondent recruitment, because we then understand why some interviewer strategies have a great impact on some respondent group's participation while other strategies have little effect. Hence, knowledge of group-specific interviewer effects is of key relevance for survey practitioners who aim to implement interviewer-related nonresponse reduction strategies.

2.2 Coding strategies for random slopes

Most investigations of interviewer effects on nonresponse use multilevel models (for a discussion of the models see Hox, 1994; Steele and Durrant, 2011). Multilevel models can account for the survey data structure, where respondents are nested in interviewers

(Durrant and Steele, 2009; Vassallo et al., 2017). Multilevel models adjust for dependencies between the levels by extended error terms (for a statistical formulation see Bryk and Raudenbush, 1992; Hox, 2010; Maas and Hox, 2004).

In the following, we highlight possible parameterization strategies for categorical grouping variables in a two-level model with random slopes. Independently of how the model is organized, we can code categorical independent variables either as contrast or separately (Verbeke and Molenberghs, 2000, ch. 12.1). When contrast coding is used for a particular independent variable, the intercept refers to the estimate for this reference category. Separate coding omits this intercept from the model so that an estimate for every category is obtained directly.

In this study, we illustrate the effect of using either of the two coding strategies in the random part of a multilevel model. For the illustration, we use data from the recruitment into the GIP. The respondent group variable to which we apply the two coding strategies is an indicator of whether a sample unit was offline or online at the time of recruitment. The paper demonstrates the differences in the statistical formulation and in the interpretation of estimates between modeling the random slopes with contrast coding and modeling them with separate coding.

The model building consists of three parts. First, we formulate a two-level model with random slopes using contrast coding. However, this parameterization does not allow us to estimate the size of interviewer effects for offliners and onliners separately. By consequence, it is difficult to draw conclusions about whether the interviewer effect for onliners differs in size from that for offliners. Therefore, we next introduce separate coding in the random part of the model to investigate differences in the size of interviewer effects per respondent groups. Finally, we include interactions of interviewer characteristics with the offliner dummy to investigate which interviewer characteristics influence interviewer effects on nonresponse differently for onliners and offliners.

2.2.1 Parameterization with contrast coding

We commence with the standard procedure typically used in the multilevel modeling of interviewer effects. We denote our dependent variable π_{ij} as a response to the online panel of respondent i who was interviewed by interviewer j. We define

$$\pi_{ij} = \begin{cases} 0 & \text{nonresponse to an online panel} \\ 1 & \text{response to an online panel.} \end{cases}$$

To estimate the between-interviewer variation in the probability to respond to the online panel we estimate a multilevel logistic regression model with two levels (respondents nested within interviewers). As in single-level logistic regressions, the probability π of observing the value 1 in the dichotomous variable π_{ij} is modeled as a logistic transformation. Resulting in

$$logit(\pi_{ij}) = ln(\frac{\pi_{ij}}{1-\pi_{ij}})$$

with π_{ij} the probability of respondent i to be approached by interviewer j responding to the online panel. Being an offliner versus being an onliner is introduced as a dummy predictor affecting the probability to respond, which is coded as

$$OFF = \begin{cases} 0 & \text{being online} \\ 1 & \text{being offline.} \end{cases}$$

Since the differences between offliners and onliners in the probability to respond can vary across interviewers, we additionally include a random slope for the offliner dummy. The resulting multilevel model is formalized in equations (2.1) to (2.3). By substituting equations (2.2) and (2.3) into (2.1), we obtain the model in reduced form in equation (2.4).

$$logit(\pi_{ij}) = \beta_{0j} + \beta_{1j}OFF_{ij} \tag{2.1}$$

$$\beta_{0j} = \gamma_{00} + u_{0j} \tag{2.2}$$

$$\beta_{1j} = \gamma_{10} + u_{1j} \tag{2.3}$$

$$logit(\pi_{ij}) = \gamma_{00} + \gamma_{10}OFF_{ij} + u_{1j}OFF_{ij} + u_{0j} \tag{2.4}$$

with

$$u_{0j} \sim N(0, \sigma^2_{u_0}), u_{1j} \sim N(0, \sigma^2_{u_1}).$$

In this model, parameters γ_{00} and γ_{10} are the fixed effects. γ_{00} is the grand intercept,

representing the logit of response for offliners across all interviewers. γ_{10} captures how the logit of response differs for offliners compared to onliners, again on average across all interviewers. The variation across interviewers is incorporated in the random part of the model. Random intercept u_{0j} denotes how the offliners' response deviates from the average for interviewer *j*. The random slope u_{1j} for being offliner represents whether the regression coefficients for onliners/offliners differ between interviewers. Therefore, the random intercept variance $\sigma^2_{u_0}$ represents the cross-interviewer variation in the success of recruiting offliners. The random slope variance $\sigma^2_{u_1}$ is less intuitive to interpret and refers to how the difference in response probabilities between offliners and onliners varies across interviewers.

In sum, in the parameterization with contrast coding, the random slope variation shows how the gap in response between offliners and onliners is different per interviewer, and thus yields insight into whether there are interviewers who "specialize" in convincing offliners or rather onliners to participate.

2.2.2 Parameterization with separate coding

However, our research questions do not focus on the interviewer effect for a *difference* in response probabilities between offliners and onliners. Instead, we are interested in the size of the interviewer effect for offliners, in the size of the interviewer effect for onliners, in whether these interviewer effects differ significantly, and in the predictors of these interviewer effects.

Thus, to answer our research questions, the parameterization of model (2.4) needs to be adapted. For this purpose, we switch from contrast coding in the fixed and random parts of the multilevel model to a model that retains contrast coding for the fixed part, but in the random part uses separate coding i.e., it introduces two dummies, one for onliners and one for offliners (see Jones, 2013, p. 136–138). The model with contrast coding in the fixed part and separate coding in the random part is formalized as

$$logit(\pi_{ij}) = \gamma_{00} + \gamma_{10}OFF_{ij} + u_{1j}OFF_{ij} + u_{2j}ON_{ij} \tag{2.5}$$

with

$$u_{1j} \sim N(0, \sigma^2_{u_1}), u_{2j} \sim N(0, \sigma^2_{u_2})$$

where

$$ON = \begin{cases} 0 & \text{being offline,} \\ 1 & \text{being online.} \end{cases}$$

This model has no random intercept (u_{0j}), but instead contains two random slopes, one u_{1j} for offliners and a second u_{2j} for onliners. However, the fixed part of equation (2.5) is identical to the fixed part of equation (2.4).

The variance components of the two random slopes reveal the size of the interviewer effect for each group separately. They thus answer our second research question as to whether the interviewer effect for onliners and differs in size from that for offliners. A significant positive covariance between the random slope for onliners and the random slope for offliners means that interviewers who are good at gaining a response from onliners are also good at gaining a response from offliners. And a significant negative covariance means that interviewers who are good at gaining a response from one group are bad at gaining a response from the other. As such, the parameterization with separate coding yield valuable insights for survey researchers.

Although both parameterizations yield different insights, it is important to stress that both models are statistically equivalent. This means that we can transform the covariance and variance from equation (2.4) into the variance component of equation (2.5) (for further details see Rabe-Hesketh and Skrondal, 2012, ch. 11.4). This statistical equivalence between models (2.4) and (2.5), however, only holds if we estimate an unstructured covariance matrix for the random effects in model (2.5) (see Rabe-Hesketh and Skrondal, 2012, ch. 11.4). Thus, we allow a correlation between both random effects at the interviewer level. The resulting variance-covariance matrix for the random slopes u_{1j}, u_{2j} is given by

$$Var \begin{bmatrix} u_{1j} \\ u_{2j} \end{bmatrix} = \begin{bmatrix} \sigma^2_{1j} & \sigma_{21} \\ \sigma_{12} & \sigma^2_{2j} \end{bmatrix}$$

To investigate group-specific interviewer effect sizes, we estimate whether the size

of the interviewer effect for onliners significantly differs from the size of the interviewer effect for offliners. For this purpose, we estimate the significance of the difference between the random slopes $u_{1j}OFF_{ij}$ and $u_{2j}ON_{ij}$ following an approach by Molenberghs and Verbeke (2007), who suggest a Wald test statistic with mixture distribution of χ^2 for two-sided hypotheses test in unconstrained multilevel models (see also Snijders and Bosker, 2011, p. 99).

2.2.3 Including cross-level interactions

The random slope model with separate coding can be augmented by including cross-level interactions between respondent groups and interviewer characteristics. Conceptually, these cross-level interactions allow us to evaluate which interviewer characteristics explain the interviewer effects among respondent groups. The extended model can be written as follows

$$logit(\pi_{ij}) = \gamma_{00} + \gamma_{10}OFF_{ij} + \gamma_{11}OFF_{ij}Z_j + \gamma_{01}Z_j + u_{1j}OFF_{ij} + u_{2j}ON_{ij} \tag{2.6}$$

with

$$u_{1j} \sim N(0, \sigma^2_{u_1}), u_{2j} \sim N(0, \sigma^2_{u_2})$$

and with $\gamma_{11}OFF_{ij}Z_j$ reflecting the intercept γ_{11} of the interaction of the respondent characteristic of being offline OFF_{ij} and the interviewer characteristics Z_j. Because we include this interaction in the fixed part of the model, we use contrast coding. This allows us to identify whether the effect of being offline on nonresponse varies as a function of a specific interviewer characteristic.

2.3 Data

This paper showcases how separate coding for the random part of an interviewer effects model can deliver insights into potentially differential interviewer effects for specific respondent groups. The case examined here investigates interviewer effects on onliners

and offliners during the recruitment of the GIP. Consequently, we use data from the GIP sample, which was recruited in 2012 and 2014 (for more details see Blom et al., 2015).

In both GIP rounds, the online panel sample was recruited in two stages. In the first stage, an interviewer-assisted face-to-face interview was conducted in a random sample of private households clustered within areas in Germany. In the second stage, all household members aged 16 to 75 in interviewed households were invited to join the online panel. We use the cases from the face-to-face interviews as our gross sample; based on these cases we model interviewer effects on response to the online panel.

We used the combined data from the 2012 and 2014 recruitment rounds. In total, 324 interviewers interviewed 5,238 age-eligible respondents during the face-to-face stage. Of these, 3,842 agreed to participate in the online panel (2,970 onliners and 872 offliners).

To account for differences in the sample composition of the interviewers' assignments, we control for respondents' age, gender, household size, level of education, employment status, the frequency of internet use, the frequency of media consumption, and whether they voted in the last general election. A small proportion of missing values on the year of a respondent's birth (<1%) was imputed using predictive mean matching (see Little, 1988; Morris et al., 2014). The analyses are presented unweighted since we aim to infer to interviewer behavior rather than the general population. Furthermore, sensitivity analyses showed no effects of the sampling design weights (which included regional clusters) on our estimates (see Blom et al., 2017).

In addition, we use data from an interviewer survey conducted during the interviewer training. This paper-and-pencil survey covers topics on interviewers' own behavior, interviewers' experience with measurements, interviewers' expectations, interviewers' computer and internet usage and interviewers' socio-demographic characteristics (interviewer survey adapted from Blom and Korbmacher, 2013). 274 interviewers completed the interviewer questionnaire.

Since the survey agency only allowed us to identify which interviewers worked on both recruitment rounds but did not allow us to match interviewers across rounds, we could not account for this additional clustering in our models. Therefore, interviewers who participated in both recruitment rounds were excluded from the analysis (57 interviewers). For eleven of the remaining 217 interviewers, a few missing values had to be imputed by predictive mean matching (<1%).

2.4 Results

To investigate interviewer effects on onliner and offliner nonresponse we follow the analytical steps set out in the methods section of this paper. For this purpose, we estimated several two-level logistic regression models (respondents nested within interviewers) with response to the GIP online panel as the dependent variable.

To estimate the size of the interviewer effect in our analyses, we commence with a null model, which only controls for respondent characteristics to account for differences in interviewer assignments. For this model (not presented) we identify an intra-class correlation of 25 percent. Thus, a considerable 25 percent of the overall variance in the GIP online panel response is located at the interviewer level.

2.4.1 Parameterization with contrast coding

Next, we estimate the same model but include a random slope for the offliner dummy (model 1 in table 2.1). To investigate whether the interviewer effect on the response of onliners and offliners differs, we estimated a log-likelihood ratio test on the model with and without random slope and tested it against a χ^2 mixture distribution. Including a random slope for offliner increases the goodness of the model significantly ($\chi^2 = 17.25$, $d.f. = 1$, $p = 0.05$). Thus, there is a significant variation between interviewers with regard to the association of offliners and onliners in response to the online panel.

This model 1 contains contrast coding for the variable being offline in both the fixed and the random parts (estimation equivalent to equation 2.4) and controls for various respondent characteristics to account for differences in interviewer assignments. We find that being an offliner significantly reduces the propensity to respond to the GIP online panel.

When we look at the random part of model 1, we find significant variance at the interviewer level ($\hat{\sigma}^2_{u_0}$) i.e., a significant interviewer effect. This means that interviewers differ in their success in recruiting respondents into the GIP online panel. In addition, there is a significant variance of the distribution of the interviewer-level slopes of being offline ($\hat{\sigma}^2_{u_1}$) i.e., there is variation between interviewers with respect to the difference in their success in recruiting onliners and offliners. We do not interpret the covariance between intercepts and slopes ($\hat{\sigma}_{u_0,u_1}$), as we cannot interpret this covariance meaning-

Table 2.1: Multilevel logistic regression of interviewer effects on response to the online panel with random slopes.

	Model 1 contrast coding		Model 2 contrast and separate coding		Model 3 cross-level interactions	
	$\hat{\beta}$	Std. err.	$\hat{\beta}$	Std. err.	$\hat{\beta}$	Std. err.
Fixed part						
Respondent characteristics						
Being offline	−0.89***	0.18	−0.89***	0.18	−3.19***	0.86
Interviewer characteristics						
Being offline and adapt to respondent					0.68**	0.25
Intercept	0.41	3.16	0.41	3.16	1.84	3.17
Random part						
$\hat{\sigma}^2_{u_1}$	0.68*	0.33	0.49*	0.22	0.42*	0.20
$\hat{\sigma}^2_{u_0}$	1.70***	0.37				
$\hat{\sigma}_{u_0,u_1}$	−0.94**	0.32				
$\hat{\sigma}^2_{u_2}$			1.71***	0.38	1.66***	0.36
$\hat{\sigma}_{u_1,u_2}$			0.76***	0.19	0.79***	0.19
Number of interviewers	214		214		214	
Number of respondents	3,751		3,751		3,751	

Note. - Three interviewers (who interviewed 64 respondents) and 27 respondents failed to respond to all questions used in the models. All models control for the respondent characteristics age, age squared, gender, household size, educational level, occupational status, internet usage, media consumption, and voting behavior. Furthermore, all models control for the interviewer characteristics age, age squared, gender, educational level, expectation of overall response rate, and adaptation to respondents.
$\hat{\beta}$=coefficients , Std. err.=standard errors
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

fully (for further explanations see Hox, 2010, p. 18).

2.4.2 Parameterization with separate coding

In the second model in table 2.1, we use contrast coding in the fixed part and separate coding in the random part (estimation equivalent to equation 2.5). As model 1 and model 2 are statistically equivalent, we do not find any differences in the coefficient estimations for the fixed part. However, the variance decomposition of the random part yields different insights.

First, we do not report an overall interviewer variance ($\hat{\sigma}^2_{u_0}$), as we omitted the random intercept for the parametrization of the separate coding. Secondly, we detect a significant random slope effect for offliners ($\hat{\sigma}^2_{u_1}$). This means that there is variation across interviewers in their success at recruiting offliners. For onliners, we also find significant interviewer effects ($\hat{\sigma}^2_{u_2}$). Interestingly, the size of the interviewer variance is considerably smaller for offliners than the interviewer variance for onliners ($\hat{\beta} = 0.49$ versus $\hat{\beta} = 1.71$). Thus, there is much less variation between interviewers when recruiting offliners compared to onliners. In addition, the difference in the size of the interviewer effects is significantly different for onliners and offliners ($\chi^2 = 8.38$, $d.f. = 2$, $p = 0.00$). Furthermore, we calculated symmetric confidence intervals for the slope variances (as suggested by Snijders and Bosker, 2011) and did not find an overlap, indicating a significant difference between the interviewer effect sizes. Finally, there is a significantly positive covariance of the two random slope coefficients ($\hat{\sigma}_{u_1,u_2}$), indicating interviewers who are good at gaining a response from onliners are also good at gaining a response from offliners and vice versa.

2.4.3 Including cross-level interactions

In the third model in table 2.1, we extend our analysis by estimating cross-level interactions (estimation equivalent to equation 2.6). Several cross-level interactions are insignificant (not presented); however, the interaction of being offline with whether interviewers say that they deviate from standardized interviewing protocols (whether they adapt to dialects and the reading speed) is significant. The positive interaction effect means that interviewers' tendency to deviate from standardized interviewing protocols

is more relevant when recruiting offliners than when recruiting onliners.

2.5 Discussion

This article investigates interviewer effects on respondents' characteristics with regard to nonresponse to an online panel. Interviewers can contribute differently to sample composition when their success in gaining response for online panel continuation differs according to respondent characteristics. Gaining knowledge on the size of interviewer effects in specific groups of respondents helps to evaluate whether an adjustment of group-specific interviewer training or other interviewer-related fieldwork strategies (e.g., matching interviewers and respondents, or targeting specific respondents with more successful interviewers) are effective for specific respondent groups. Therefore, we investigated the interviewer effect size for different groups of respondents using a multilevel logistic regression with an alternative parametrization strategy for respondent groups within the random slopes, namely separate coding.

Separate coding allows us to answer the question of whether the interviewer effect size is the same for different respondent groups. The interviewer effect size is relevant for an evaluation of whether interviewers are the reason for low response rates among specific respondent groups and whether interviewer-related fieldwork strategies are effective for specific respondent groups. Therefore, we investigated whether the size of interviewer effects on nonresponse differs for onliners and offliners and if so, which interviewer characteristics influence these interviewer effects.

The variance at the interviewer level indicates that interviewers affect nonresponse to online panel continuation as well as different interviewer effects for onliners compared to offliners. However, the interviewer effect when recruiting offliners is about three times smaller compared to onliners, indicating that the interviewer effect on nonresponse is much larger for onliners compared to offliners. Furthermore, we find that the adaptation of pace of speech and dialect to the respondent is more relevant when obtaining a response from offliners than when obtaining a response from onliners. Interviewer training could build on this knowledge by making interviewers aware that they are more likely to perceive a response from offliners when interviewers deviate from standardized interviewing protocols (e.g., by adapting to dialects).

2.6 Conclusions

Based on our results, we conclude that interviewers have an impact on response to online panel continuation; however, the low response rates of offliners are not associated with interviewer effects indicated by the small interviewer effect size of offliners. Consequently, it is not the interviewers who introduced low response rates for offliners in the online panel sample. The small interviewer effect sizes mean that implementing or adjusting interviewer strategies for reducing nonresponse, for example by matching or targeting respondents with specific interviewers, shall not have a huge influence on response rates of offliners. Furthermore, our results suggest that the adaptation of interviewer-related fieldwork strategies might help to increase response rates among onliners, as interviewer effect size was relatively large compared to the interviewer effect size of offliners.

We found that all interviewers are equally good or bad in recruiting offliners, indicating that there is a general issue in the recruitment of offliners. This general issue in recruiting offliners could be tackled in the interviewer training. For example, interviewers might have used internet devices as an additional incentive to motivate offliners. However, this argument might not have worked as offliners consider the usage of devices as a burden. Thus, interviewers could be trained to motivate offliners more with regard to respondents' contribution to research. If a change in the interviewer training does not increase response rates of offliners, one has to find other recruitment strategies for reluctant offliners which might not involve interviewers (e.g., using a mixed-mode approach).

In addition, we found the cross-level interaction of respondents' being offline and interviewers adapting to respondents during the interview to be significant. This significant interaction indicates that gaining a response from offliners might increase compared to onliners when interviewers adapt in terms of rate of speaking or adjusting to the dialect of the respondents. However, the small interviewer effect variation of offliners indicates that a possible increase in response rates will be rather marginal.

In general, separate coding can be useful for survey practitioners and researchers, who find high or low response rates among specific respondent groups in interviewer assisted surveys. Separate coding allows an evaluation of interviewer effect size in each respondent group indicating whether interviewer effects differ in size by groups of

respondents. For example, liking (matching interviewers and respondents with sharing attributes) might be differently effective for specific groups of respondents, which can be investigated by the size of interviewer effects in advance.

Future research should investigate other respondent groups with higher or low response rates. For example, Helmschrott and Martin (2014) found high nonresponse rates among lower educated respondents in a survey of adult skills. In this context, the question at hand is, whether interviewers differ in gaining a response from lower educated respondents compared to higher and medium educated respondents and whether interviewer strategies can change the imbalances in response rates. Clearly, surveys with large imbalances among respondent groups gain from an investigation of the variation of interviewer effects. By considering the interviewer effect size, one might be more effective when implementing or adjusting interviewer-related fieldwork strategies and thus, one might mitigate nonresponse bias more effectively.

Acknowledgments

The authors disclosed receipt of the following financial support for this research, authorship, and/or publication of this article: The German Internet Panel is the central data collection project (Z1) of Collaborative Research Center 884 "Political Economy of Reforms" (SFB 884) at the University of Mannheim and is funded by the German Research Foundation (DFG).

References

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., Herzing, J. M. E., Cornesse, C., Sakshaug, J. W., Krieger, U., and Bossert, D. (2017). Does the Recruitment of Offline Households Increase the Sample Repres-

entativeness of Probability-Based Online Panels? Evidence from the German Internet Panel. *Social Science Computer Review*, 35(4):498–520.

Blom, A. G. and Korbmacher, J. M. (2013). Measuring Interviewer Characteristics Pertinent to Social Surveys: A Conceptual Framework. *Survey Methods: Insights from the Field*. https://doi.org/10.13094/SMIF-2013-00001.

Bryk, A. S. and Raudenbush, S. W. (1992). *Hierarchical Linear Models, Applications and Data Analysis Methods*. Sage, Thousand Oaks, CA.

Couper, M. P. and Groves, R. M. (1992). The Role of the Interviewer in Survey Participation. *Survey Methodology*, 18(2):263–278.

de Heer, W. and de Leeuw, E. D. (2002). Trends in Household Survey Nonresponse: A Longitudinal and International Comparison. In Groves, R. M., Dillman, D. A., Eltinge, J. L., and Little, R. J., editors, *Survey Nonresponse*, chapter 3, pages 41–54. Wiley, New York, NY.

Durrant, G. B., Groves, R. M., Staetsky, L., and Steele, F. (2010). Effects of Interviewer Attitudes and Behaviors on Refusal in Household Surveys. *Public Opinion Quarterly*, 74(1):1–36.

Durrant, G. B. and Steele, F. (2009). Multilevel Modelling of Refusal and Non-Contact in Household Surveys: Evidence from Six UK Government Surveys. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 172(2):361–381.

Groves, R. M. (2006). Nonresponse Rates and Nonresponse Bias in Household Surveys. *Public Opinion Quarterly*, 70(5):646–675.

Heerwegh, D. (2009). Mode Differences Between Face-to-face and Web Surveys: An Experimental Investigation of Data Quality and Social Desirability Effects. *International Journal of Public Opinion Research*, 21(1):111–121.

Helmschrott, S. and Martin, S. (2014). Nonresponse in PIAAC Germany. *methods, data, analyses*, 8(2):244–266.

Hox, J. J. (1994). Hierarchical Regression Models for Interviewer and Respondent Effects. *Sociological Methods & Research*, 22(3):300–318.

Hox, J. J. (2010). *Multilevel Analysis: Techniques and Applications*. Routledge, New York, NY, 2nd edition.

Hox, J. J. and de Leeuw, E. D. (2002). The Influence of Interviewers' Attitude and Behavior on Household Survey Nonresponse: An International Comparison. In Groves, R. M., Dillman, D. A., Eltinge, J. L., and Little, R. J. A., editors, *Survey Nonresponse*, chapter 7, pages 103–120. Wiley, New York, NY.

Jones, K. (2013). *Developing Multilevel Models for Analysing Contextuality, Heterogeneity and Change Using MLwiN 2.2*, volume 1. SV Subramanian, Bristol, UK.

Little, R. J. A. (1988). Missing-Data Adjustments in Large Surveys. *Journal of Business & Economic Statistics*, 6(3):287–296.

Loosveldt, G. and Beullens, K. (2014). A Procedure to Assess Interviewer Effects on Nonresponse Bias. *SAGE Open*, 4(1):1–12.

Maas, C. J. M. and Hox, J. J. (2004). Robustness Issues in Multilevel Regression Analysis. *Statistica Neerlandica*, 58(2):127–137.

McCabe, S. E. and West, B. T. (2016). Selective Nonresponse Bias in Population-Based Survey Estimates of Drug Use Behaviors in the United States. *Social Psychiatry and Psychiatric Epidemiology*, 51(1):141–153.

Molenberghs, G. and Verbeke, G. (2007). Likelihood Ratio, Score, and Wald Tests in a Constrained Parameter Space. *The American Statistician*, 61(1):22–27.

Morris, T. P., White, I. R., and Royston, P. (2014). Tuning Multiple Imputation by Predictive Mean Matching and Local Residual Draws. *BMC Medical Research Methodology*, 14(75):1–13.

Peytchev, A., Riley, S., Rosen, J., Murphy, J., and Lindblad, M. (2010). Reduction of Nonresponse Bias Through Case Prioritization. *Survey Research Methods*, 4(1):21–29.

Pickery, J. and Loosveldt, G. (2002). A Multilevel Multinomial Analysis of Interviewer Effects on Various Components of Unit Nonresponse. *Quality and Quantity*, 36(4):427–437.

Pickery, J. and Loosveldt, G. (2004). A Simultaneous Analysis of Interviewer Effects on Various Data Quality Indicators With Identification of Exceptional Interviewers. *Journal of Official Statistics*, 20(1):77.

Rabe-Hesketh, S. and Skrondal, A. (2012). *Multilevel and Longitudinal Modeling Using Stata*. Stata Press, College Station, TX.

Snijders, T. A. B. and Bosker, R. J. (2011). *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling*. Sage, London, UK.

Steele, F. and Durrant, G. B. (2011). Alternative Approaches to Multilevel Modelling of Survey Non-Contact and Refusal. *International Statistical Review*, 79(1):70–91.

Vassallo, R., Durrant, G., and Smith, P. (2017). Separating Interviewer and Area Effects by Using a Cross-Classified Multilevel Logistic Model: Simulation Findings and Implications for Survey Designs. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 180(2):531–550.

Verbeke, G. and Molenberghs, G. (2000). *Inference for the Marginal Model*, chapter 6, pages 55–76. Springer Series in Statistics. Springer, New York, NY.

West, B. T. and Blom, A. G. (2016). Explaining Interviewer Effects: A Research Synthesis. *Journal of Survey Statistics and Methodology*, 5(2):175–211.

West, B. T., Kreuter, F., and Jaenichen, U. (2013). "Interviewer" Effects in Face-to-Face Surveys: A Function of Sampling, Measurement Error, or Nonresponse? *Journal of Offical Statistics*, 29(2):277–297.

West, B. T. and Olson, K. (2010). How Much of Interviewer Variance is Really Nonresponse Error Variance? *Public Opinion Quarterly*, 74(5):1004–1026.

Chapter 3

Paper II

The influence of a person's IT literacy on unit nonresponse and attrition in an online panel

Abstract

Research has shown that the non-internet population is hesitant to respond to online survey requests. However, subgroups in the internet population with low IT literacy may hesitate to respond to online surveys. This latter issue has not yet received much attention by scholars, despite its potentially detrimental effects on the external validity of online survey data. In this paper, we explore the extent to which a persons' IT literacy contributes to nonresponse bias in the German Internet Panel (GIP), a probability-based online panel of the general population. With a multi-dimensional classification of IT literacy, we predict response to the first online panel wave and participation across panel waves. We find that persons that belong to different classes of IT literacy have systematically different socio-demographic characteristics and show different voting behavior. In addition, we find that initial response propensities vary by classes of IT literacy, as do attrition patterns over time. Our results demonstrate the importance of IT literacy for the reduction of nonresponse bias during fieldwork and post-survey adjustments.

The influence of a person's IT literacy on unit nonresponse and attrition in an online panel[1]

3.1 Introduction

Online surveys have become a prevalent data source for market research and behavioral sciences (Baker et al., 2010, p. 7; Schonlau et al., 2002). The online mode is attractive because it enables conducting interviews cost-effectively in terms of time, space, and labor (Greenlaw and Brown-Welty, 2009; Hardigan et al., 2012; Kaplowitz et al., 2004). Furthermore, research finds low measurement error in online surveys due to a reduction in social desirability bias (Kreuter et al., 2008) and cancellation of interviewer effects in the self-completion mode (for a review see West and Blom, 2016).

However, researchers have raised concerns about the generalizability to the general population of estimates based on online surveys (Best et al., 2001; Bethlehem, 2010; Dever et al., 2008; Mohorko et al., 2013; Sterrett et al., 2017). Much of this discussion circles around the suitability of nonprobability sampling methods employed by many commercial online survey providers (Bethlehem and Stoop, 2007; Gelman et al., 2016; Yeager et al., 2011).

Because of apparent selectivities in nonprobability online surveys, there is now an increasing number of probability-based online panels, which acknowledge the need for both probability sampling and coverage of persons without computers and/or internet (see for example Blom et al., 2017; Bosnjak et al., 2013; de Vos, 2010; Knoef and de Vos, 2009; Revilla et al., 2016). Most probability-based online panels account for potential coverage biases by either equipping so-called offliners with devices and internet or by interviewing them via a different mode such as mail questionnaires (see Blom et al., 2016; Bosnjak et al., 2017). However, even when covered, offliners tend to be underrepresented in the online sample, because they are less likely to respond to the request to become part of the online panel (Blom et al., 2017; Hoogendoorn and Daalmans, 2009; Leenheer and Scherpenzeel, 2013). Research further shows that offliners differ from onliners in their socio-demographic characteristics (Blom et al., 2017; Leenheer and Scherpenzeel, 2013), their attitudes and behavior (Zhang et al., 2008), and their

[1] This chapter is co-authored with Annelies G. Blom and is currently under review.

general health (Schnell et al., 2017). Nevertheless, online surveys seem to become the dominant data source for political, sociological, and economic research. Therefore, we need to understand the mechanisms of selectivities and discover ways to adjust for them to avoid biased estimates in substantive research.

Studies of selectivities in probability-based online samples have thus far considered the use of information technologies (IT) a binary phenomenon: persons were considered to be either offline or online. However, there has been a change in the research on the digital divide, away from a mere digital access divide towards a digital usage divide and digital device divide (Antoun, 2015; van Deursen and van Dijk, 2014). This change in the digital divide is of particular interest to survey research because it is likely associated with response to online surveys (for the association between response rates and mode preferences see Millar and Dillman, 2011; Olson et al., 2012; Rookey et al., 2008; Shih and Fan, 2007).

This change in the terminology in the digital divide literature is due to increasing internet penetration rates and people participating in digital developments to different extents resulting in differences in device usage between people. For example, the European Commission (2014, p. 44) reports that in 2014 46 percent of German households had a mobile and home internet access, 30 percent of German households had only home internet access, 3 percent had only mobile internet access, whereas 21 percent had no internet access at all. Since there are large differences in internet access and hence device usage, we can assume that persons respond differently to online survey requests, depending on whether they have access to a computer, tablet, smartphone, and internet connection. The notion that the differences in internet access and device usage affect response to the online survey requests is supported by Barron and Dennis (2016) and Pforr and Dannwolf (2017), who find that about one fourth of those who use the internet for private purposes prefer a paper questionnaire or telephone interview over an online questionnaire in mixed-mode surveys. Thus, it is not only offliners who hesitate to respond to an online panel request but also many onliners. Therefore, we need a more fine-grained concept of IT usage, to account for further subgroups in the online population (for different concepts of IT literacy see Blank, 2016; Blank and Groselj, 2014; Brandtzæg, 2010; Dutton and Blank, 2014; van Deursen and van Dijk, 2014; Wang et al., 2013). In addition to drawing attention to the concept of IT literacy, we aim to find a *single* predictor of unit nonresponse in online surveys, because

this enables the timely and efficient development of adjustments for unit non response both during fieldwork (for example for case prioritization and in adaptive or responsive designs Brick and Tourangeau, 2017; Kreuter et al., 2010; Peytchev et al., 2010) and for weighting and imputation purposes post-hoc. The fruitful combination of measuring IT literacy and investigating unit nonresponse in an online panel allows us to add to current research into unit nonresponse and nonresponse bias in probability-based online panels.

In the following, we develop the multi-dimensional measure of IT literacy from a theoretical argument and investigate IT literacy empirically by analyzing data from the face-to-face recruitment interview of the German Internet Panel (GIP). By using Latent Class Analysis (LCA), we gain knowledge about the patterns in a set of sample person characteristics and from these extract classes of sample persons. We then investigate the added value of these classes of IT literacy in explaining the response to both the first online panel wave and long-term attrition in the GIP.

3.2 Theoretical and conceptual framework

3.2.1 IT literacy

Traditionally, IT literacy relates to the concept of the digital divide, which is based on a dichotomy between "haves" and "have-nots" regarding the means to connect to the internet (DiMaggio et al., 2004, p. 2; Rice and Katz, 2003). Since internet penetration rates and the variety of internet devices available are increasing in Western countries (European Commission, 2016; Mohorko et al., 2013; Sterrett et al., 2017), the term digital divide is ever changing, resulting in a variety of different definitions which cover different aspects of access to and competence in the use of IT and/or the internet (see Guo et al., 2008; Wang et al., 2013).

The digital divide literature increasingly argues that the digital access divide turns into a digital usage divide (for example van Deursen and van Dijk, 2014). Research investigating the digital usage divide explores the characteristics and quality of typologies regarding IT usage (for an overview see Blank and Groselj, 2014; Brandtzæg et al., 2011; Brandtzæg, 2010; van Deursen and van Dijk, 2014). In this respect, the literature on IT literacy typologies identifies three dimensions of internet usage (see Blank and Groselj, 2014; Brandtzæg, 2010; Holmes, 2011): (1) amount or frequency of internet

use, (2) variety of internet use, and (3) type or content of internet use. The amount of internet use is measured either by the frequency of internet use or by the length of time someone uses the internet (Brandtzæg et al., 2011; Brandtzæg, 2010). The variety of internet use is operationalized by the number of different activities persons are engaged in, such as the use of Twitter, Facebook, or LinkedIn (Holmes, 2011). The type of internet usage reflects content preferences. For example Blank and Groselj (2014) distinguish entertainment, commerce, information seeking, socializing, e-mailing, blogging, schoolwork, classic mass media usage, and others. In addition, Antoun (2015) extends the concepts of digital access divide – being online or offline – and digital usage divide by adding the concept of digital device divide. Antoun (2015) identifies four categories of survey respondents: non-internet users, computer internet users, mobile internet users, internet users that use a combination of mobile devices and computers. Furthermore, we know from decision models regarding technology adaptation that initial and continued IT usage depends on the perceived usefulness of and satisfaction with IT (see Hong et al., 2006). For example, Porter and Donthu (2006) report that even when controlling for socio-demographic differences and IT access barriers, perceptions regarding the use and usefulness of the internet have a strong effect on its usage. This process may well translate to online surveys such that attitudes towards IT may impact on persons' likelihood to participate in an online panel.

Based on this literature, researchers investigating nonresponse in probability-based online panels may be well-advised to consider a multidimensional classification of IT literacy that includes aspects of the digital access divide, the digital device divide, the digital usage divide, as well as attitudes towards technical innovation.

3.2.2 Hypotheses

IT literacy may be a valuable predictor of nonresponse bias and useful auxiliary variable for post-adjustment procedures if it is associated with both key survey variables and sample units' propensity to respond (see Groves' (2006) common cause model).

First, we consider the *association between IT literacy and key survey variables*. Previous research has shown that internet access and usage vary according to age and educational level (for examples see Antoun, 2015; de Bruijne and Wijnant, 2014; Dutton and Blank, 2014; Friemel, 2016; Rice and Katz, 2003; Slegers et al., 2012; van Deursen

and van Dijk, 2014, 2015). Furthermore, research has reported differences in news consumption between internet users and nonusers, and within subgroups of internet users (Stempel et al., 2000; van Deursen and van Dijk, 2014). Moreover, Robinson et al. (2002) and Wei and Hindman (2011) found an association between internet usage and political opinion. Based on this literature, we derive the following hypothesis regarding the association between IT literacy and key survey variables.

Hypothesis 1 *IT literacy differs by age, educational level, media consumption, political interest, and voting behavior.*

When, we consider general predictors of unit nonresponse bias, the decision to participate in surveys is related to a person's predispositions, such as their socio-demographic characteristics (Couper et al., 2007; Hoogendoorn and Daalmans, 2009; Kaplowitz et al., 2004; Payne and Barnfather, 2012; Rookey et al., 2008) and the survey topic (Groves et al., 2004; Zillmann et al., 2014). We expect a persons' IT literacy to be an *additional factor in explaining unit nonresponse bias* in probability-based online panels. In accordance with cost-benefit theory, we expect that persons differentially perceive the benefits and costs of participating in an online panel (for an elaboration see Schnell, 1997, p. 133–216; Singer, 2011, p. 381). Due to differences in their IT literacy, some persons will experience response to an online panel as more burdensome than others (see Fuchs and Busse, 2009; Hoogendoorn and Daalmans, 2009; Kwak and Radler, 2002). And how burdensome persons consider their response to an online panel will influence the subjective costs in their subconscious cost-benefit calculation. As described in the unit nonresponse literature respondents face various costs when participating in a survey (for example Singer, 2011). For the sake of simplicity, we assume that other factors that may influence a sample unit's costs of participation (i.e., those not related to IT literacy) are evenly distributed across different classes of IT literacy.

Hypothesis 2 *IT literacy is an additional predictor of unit nonresponse in an online panels.*

Based on this hypothesis we argue that the *costs of participating in an online panel* are lowest for people who regularly use the internet and highest for people without a computer and/or internet access (see also Kwak and Radler, 2002). Therefore, we

expect that, for probability-based online panels, persons' response propensity will vary according to their class of IT literacy.

Hypothesis 3 *IT literacy groups differ in their propensity to respond to an online panel requests.*

Once a person becomes a member of an online panel, their costs of continued response to online panel waves may change over time, because they receive access to and gain experience with filling out the online questionnaires (for different types of attrition patterns see Lugtig, 2014). For example, Leenheer and Scherpenzeel (2013) find that non-internet households are more loyal panel participants than internet households. In this context, it is conceivable that, once previously offline persons are online, their response costs are reduced, resulting in lower attrition rates over time. For persons using computers and the internet regularly, in contrast, we do not expect the response costs to change, because they do not gain any significantly new IT skills through their online panel participation.

Hypothesis 4 *IT literacy groups differ in their propensity to continue participating in online panel waves.*

In the following, we explore these hypotheses one-by-one and thus draw a comprehensive picture of the influence of IT literacy on unit nonresponse and unit nonresponse bias in probability-based online panels.

3.3 Data

To investigate the contribution of IT literacy to explaining unit nonresponse bias, we use data from the German Internet Panel (GIP). The GIP is a probability-based online panel of the general population aged 16-75, for which respondents are initially recruited via face-to-face surveys and subsequently invited to the online panel. This procedure ensures that (a) the gross sample can be drawn with strict probability methods and (b) the gross sample includes the offline population. Offliners are enabled to participate in the panel by providing them with the necessary IT equipment and support. GIP panel members are invited to participate in online surveys on a variety of political, social, and economic issues every two months (see Blom et al., 2015).

The GIP recruitment data offers a unique opportunity to gain insights into the non-response processes in an online panel. Since the GIP is based on a random sample of the general population and was recruited entirely offline, its recruitment data hold rich information on different aspects of IT literacy, independent of the sample persons' eventual participation in the online panel.

We use data from the 2014 face-to-face recruitment interview and, based on this data, model response to the survey. The GIP is sampled and recruited in multiple stages (see table 3.1). First, 299 areas were randomly selected (for more details see Blom et al., 2015). Next, 9,316 eligible households were randomly selected within the 299 areas. Finally, 4,426 face-to-face interviews were conducted with one person per eligible household. Out of the 4,426 persons interviewed face-to-face 3,985 people were identified as age-eligible, i.e., born in the years 1938 to 1997. In our models, these constitute the gross sample, and we refer to them as *sample units or sample persons*. Of these 3,985 sample units, 2,064 participated in the first online panel wave. In the following, we refer to them as respondents.

Table 3.1: GIP data structure.

Recruitment process		n
regional clusters		299
eligible households		9,316
individuals interviewed face-to-face		4,426
eligible individuals interviewed face-to-face	sample units	3,985
eligible individuals participated online	respondents	2,064

NOTE. – n = number of observations.

Within each randomly selected household, exactly one person willing to do the interview was interviewed face-to-face. Overall, this resulted in unequal selection probabilities at the face-to-face interview. To account for this, we apply design weights (for a description see Blom et al., 2017). In addition, we account for regional clusters in the sampling design by means of jackknife variance estimation (for details see Gould, 1995; Quenouille, 1956). A small proportion of item missing information on the variable age ($< 0.5\%$ refused to answer) was imputed using predictive mean matching (for details see Little, 1988; Morris et al., 2014).

3.4 Methods

Following Holmes' (2011) typological model of IT literacy, we investigate our theoretical argument for a multi-dimensional conceptualization of IT literacy with a Latent Class Analysis (LCA). The LC framework categorizes sample units into classes based on their similarity in answer patterns. The latent classes are parameterized by means of a maximum-likelihood estimation. Sample units are assigned to a latent class based on their highest class probability in terms of their conditional probabilities for values of the predictor variables. Accordingly, sample units who hold similar characteristics in terms of IT access and usage have a high probability of belonging to the same class of IT literacy.

To select the best fitting LC model we use six evaluation criteria (McCutcheon, 2002; Nylund et al., 2007; Vermunt and Magidson, 2016): the number of parameters, the degrees of freedom, the p-value of the χ^2 statistic, the change in the log-likelihood (LL), the Bayesian Information Criterion (BIC), and the likelihood ratio χ^2 statistic (L^2). To formally assess the number of IT literacy classes prevalent in our data, we vary the number of classes in the LCA step-by-step from 2 to 6[2] and compute our six evaluation criteria (for results see below). Models with more than 6 classes did not show any significant changes in the evaluation criteria.

Having selected the model according to these criteria, we investigate whether the detected IT literacy classes contribute to explaining nonresponse bias in the online panel. We test whether sample units are belonging to different IT literacy classes also vary in key socio-demographic characteristics and political behavior. For this purpose, we conduct tests of differences, such as mean-comparisons and Pearson's χ^2 statistics. Next, to examine whether general predictors of nonresponse correlate with sample units' propensity to respond, we estimate logistic regression models in which we regress response to the first online panel wave on the general predictors of nonresponse age, gender, educational level, household size, urbanity, political interest, and electoral participation. Subsequently, to investigate the additional contribution of IT literacy, we extend the analysis by our classes of IT literacy and perform a log-likelihood ratio test (LR test). In the final model, we extend the analysis by interaction terms between IT

[2] We tested more than 6 classes in the LCA; however we do not gain more information as the evaluation criteria do not change significantly compared to the 6 class model.

literacy and general predictors to examine possible interdependencies of IT literacy with other variables.

Finally, we explore the added value of IT literacy in explaining attrition in the GIP. For this purpose, we run a logistic regression model with a binary dependent variable of participation in each panel wave on the pooled data of the GIP. We treated the ten panel waves as a continuous variable and included panel waves as a quadratic term, based on the theoretical argument that respondents' costs of participation diminish from the first online questionnaire to each additional online questionnaire, as they become more familiar with the survey process (the empirical nonlinearity was also graphically tested). In a second step, we extended the model by interaction terms between waves and classes of IT literacy to investigate separate time-trends for each class of IT literacy.

3.5 Operationalization

To identify peoples' IT literacy with an LCA we derived measures of digital devices, digital access, digital usage, and attitudes towards technology from the GIP face-to-face recruitment interview. First, a sample unit's digital devices and access were derived from a combination of questions on access to a desktop, laptop or tablet computer, access to the internet, and ownership of a smartphone. From this, we derived two variables: an indicator of whether a sample unit has a computer at home, a computer and internet at home, or neither a computer nor internet at home, and an indicator of whether a sample unit owns a smartphone. Second, digital usage was identified with a question on a sample unit's frequency of internet use ("How often do you use the internet, the world wide web or e-mail for private purposes, whether at home or work?"). Finally, a question evaluates sample units' attitudes towards technical innovations ("How important is it for you that your technical equipment at home, like mobile phone, television or computer, is cutting edge technology?").

3.6 Results

3.6.1 Measuring IT literacy

Table 3.2 shows the six model evaluation criteria (number of parameters, degrees of freedom, the p-value of the χ^2 statistic, change in the log-likelihood (LL), BIC, and the likelihood ratio χ^2 statistic (L^2)) for a series of LCA models with different numbers of classes.

Table 3.2: Goodness of fit for 2-6 classes.

# classes	# parameters	d.f.	p-value	LL	BIC	L^2
2	14	81	0.00	-13,257	26,631	716
3	19	76	0.00	-12,990	26,138	181
4	24	71	0.00	-12,954	26,108	110
5	29	66	0.01	-12,946	26,134	94
6	34	61	0.09	-12,938	26,157	77
Base 3,979						

NOTE. – All models were estimated using LatentGOLD (Vermunt and Magidson, 2016). Six sample units failed to respond to all questions used for the LCA and were case-wise deleted.
d.f.= degrees of freedom
p-value = p-value of χ^2
LL = log-likelihood
L^2 = likelihood ratio of χ^2

The 4-class model shows the best overall results for the statistics examined: a low p-value, the lowest BIC, a relatively low L^2, and LL. With an associated p-value of 0.00, the 4-class solution fits the data well. The BIC value reaches a minimum at the 4-class solution, but is higher for 2- or 3- and 5- or 6-classes, indicating that we can indeed identify four subgroups with distinctive characteristics. Furthermore, when moving from 3- to 4- and 4- to 5-classes, the LL fails to decrease substantially. Therefore, we select the model with the lowest BIC. In addition, the L^2 (110) for the 4-class model is not substantially larger than the respective degrees of freedom (71), indicating a good model fit. These results are robust to sample size, as we estimated the analysis twice, once with the full dataset and once with a 90 percent random sub-sample. The assumption of local independence was not violated for the 4-class model. Consequently, the

statistical arguments suggest a 4-class solution for the latent construct IT literacy.

To allow a closer look at the composition of each IT literacy class, table 3.3 presents conditional probabilities of all variables used in the LCA for the 4-class model. Based on this composition, we describe the four classes as follows:

- *Digital addicts* have a computer and internet at home (0.96) and use new technological devices, like smartphones (0.77). They use the internet on a daily basis (0.96), and technology is important to them (very important 0.23 and important 0.44). Thus, this class of persons is motivated and able to use the internet on a daily basis and has access to smartphones.

- *Late adopters* have a computer and internet at home (0.90), and some of them own a smartphone (0.32). About one-third of the late adopters use the internet on a daily basis (0.34). On the whole, this class of persons attaches little importance to technology (very important 0.06 and important 0.29). Thus, late adopters have access to the internet, and some use a smartphone; however, their lives do not revolve around IT.

- *Traditional users* often have computers, but no internet at home (0.51); however, some of them do not have computers and internet at all (0.44). This class of persons does not own a smartphone (0.00), although half of them use the internet on a daily basis (0.54). Furthermore, technology is not important to persons belonging to this class (very important 0.11 and important 0.36). Consequently, the majority of this class is familiar with the internet via computers (maybe in their workplace), but not with smartphones.

- *Digitally excluded* persons do not have a computer or internet at home (0.91). They never use the internet (0.97) or only a few times a year (0.03). Furthermore, digitally excluded persons do not own a smartphone (0.00) and consider technology only slightly or not at all important (very important 0.02 and important 0.19). All in all, digitally excluded persons are bypassed by recent technological developments.

The penultimate row of table 3.3 shows the probability of sample units being in each class, which is equivalent to the size of a class. It illustrates that while 51 percent

Table 3.3: Conditional probabilities for values of the predictor variables for the 4-class model.

	Digital addicts	Late adopters	Traditional users	Digitally excluded
Equipment status				
Has computer and internet	0.96	0.90	0.04	0.00
Has computer but no internet	0.04	0.10	0.51	0.09
Has no computer and no internet	0.00	0.00	0.44	0.91
Smartphone usage				
No	0.23	0.68	1.00	1.00
Yes	0.77	0.32	0.00	0.00
Internet usage				
Never	0.00	0.11	0.04	0.97
<1 a month to once a week	0.00	0.24	0.14	0.03
>1 a week	0.04	0.30	0.29	0.00
Daily	0.96	0.34	0.54	0.00
Importance of technology				
Not important	0.04	0.17	0.09	0.28
Slightly important	0.29	0.49	0.42	0.51
Important	0.44	0.29	0.36	0.19
Very important	0.23	0.06	0.11	0.02
Class size	0.51	0.34	0.04	0.12
Base	1,944	1,398	121	516

of persons were identified as digital addicts, a minority of 34 percent of persons were classified as late adopters, 4 percent as traditional users, and 12 percent as digitally excluded.

To investigate how the LCA classification of IT literacy in this paper compares to the binary online/offline classification in previous research into nonresponse in probability-based online panels (see Blom et al., 2017; Leenheer and Scherpenzeel, 2013) we cross-tabulate respondents' membership in our IT literacy classes with the binary online/offline indicator (see table 3.4). We find that Blom et al. (2017) and Leenheer and Scherpenzeel (2013) would have predominantly defined traditional users and digitally excluded as offliners, whereas digital addicts and late adopters would have mainly been classified as onliners. Consequently, our classification of IT literacy subdivides both groups – offliners and onliners – into two subgroups.

Table 3.4: Comparison of binary and multi-dimensional classification.

	Digital addicts	Late adopters	Traditional users	Digitally excluded
Offliner	0.05	0.11	1.00	1.00
Onliner	0.95	0.89	0.00	0.00
Base 3,979				

NOTE. – Table reports proportions. In the case of digital addicts and late adopters we are observing the effects of misclassification errors associated with the assignment of sample units to a latent class based upon the models highest class probability.

3.6.2 Predictors of unit nonresponse and nonresponse bias

Having derived four classes of IT literacy from information on their access to, use of, and attitudes towards using IT, we next investigate whether sample units that belong to different classes of IT literacy also differ in their socio-demographic characteristics, key political attitudes and behavior, and response propensity. For this purpose, table 3.5 reports differences in sample unit characteristics by IT literacy. We find significant differences in the composition of IT literacy classes regarding all characteristics assessed.

As table 3.5 shows, the mean age of sample units differs significantly across the four classes of IT literacy ($F_{3;3,979} = 364.86$, $p = 0.00$). Sample units belonging to the classes late adopters and traditional users are of similar mean age (51 and 47 years,

Table 3.5: Differences in sample units' characteristics by IT literacy.

	Digital addicts	Late adopters	Traditional users	Digitally excluded	Test of difference
Mean age	41	51	47	63	***[1]
Educational level					
Low educational level	0.16	0.29	0.32	0.69	
Medium educational level	0.30	0.38	0.40	0.25	***[2]
High educational level	0.50	0.31	0.28	0.06	
Current pupil	0.04	0.01	0.00	0.00	
Political interest					
Low interest	0.45	0.41	0.36	0.40	
Some interest	0.44	0.47	0.50	0.43	***[2]
High interest	0.10	0.12	0.13	0.17	
Media consumption per day					
Never	0.05	0.06	0.09	0.09	
$< \frac{1}{2}$ hour	0.37	0.38	0.36	0.33	
$> \frac{1}{2} - 1$ hour	0.39	0.37	0.35	0.33	
$> 1 - 1\frac{1}{2}$ hours	0.10	0.11	0.11	0.14	**[2]
$> 1\frac{1}{2} - 2$ hours	0.04	0.04	0.05	0.05	
$> 2 - 2\frac{1}{2}$ hours	0.02	0.02	0.01	0.03	
$> 2\frac{1}{2} - 3$ hours	0.01	0.01	0.01	0.01	
> 3 hours	0.02	0.01	0.04	0.02	
Vote 2013					
CDU/CSU	0.20	0.21	0.16	0.27	
SPD	0.17	0.18	0.22	0.17	
The Left	0.04	0.04	0.05	0.04	
The Greens	0.11	0.08	0.07	0.02	
Other party	0.09	0.07	0.10	0.05	***[2]
Nonvoters	0.16	0.17	0.15	0.23	
Not eligible	0.09	0.06	0.12	0.05	
Don't know	0.06	0.07	0.05	0.06	
Refused to answer	0.08	0.12	0.09	0.11	
Base[3]	1,944	1,398	121	516	

NOTE. – Analyses conducted with design weights to account for unequal selection probabilities across households of different sizes. We report proportions, with the exception of mean age.
[1] p-value based on a one-way analysis of variance and covariance.
[2] p-value based on Pearsons's χ^2 statistic.
[3] Base for educational level = 3,972, political interest = 3,977, media consumption per day = 3,972.
* p < 0.05, ** p < 0.01, *** p < 0.001

respectively), whereas digital addicts are on average 41 years and digitally excluded 63 years old.

In addition, we find that IT literacy is significantly related to the educational level ($\chi^2 = 660.45$, $d.f. = 9$, $p = 0.00$). For example, 50 percent of the digital addicts compared to 6 percent of the digitally excluded have a high level of education, whereas the IT literacy classes late adopters and traditional users resemble each other in terms of their level of education (31 and 28 percent, respectively).

We also find differences between the classes of IT literacy with regard to media consumption ($\chi^2 = 38.08$, $d.f. = 21$, $p = 0.01$). For example, 9 percent of the digitally excluded and of the traditional users indicated not to use the internet, radio, television or newspapers to gather information on political issues as compared to digital addicts and late adopters with 5 and 6 percent, respectively.

Examining the attitudinal variables, we find significant differences across IT literacy classes regarding political interest ($\chi^2 = 35.63$, $d.f. = 6$, $p = 0.00$). Digitally excluded persons are most interested in politics, whereas digital addicts are least interested in politics (17 and 10 percent, respectively).

Furthermore, our results indicate that persons belonging to different IT literacy classes differ in their alleged vote choice ($\chi^2 = 102.05$, $d.f. = 24$, $p = 0.00$). The digitally excluded (27 percent) are most likely to say that they voted for the CDU/CSU (Christian Democrats) at the last general election, while only 16 percent of the traditional users, 21 percent of the late adopters, and 20 percent of the digital addicts allegedly voted for this party. In addition, the digitally excluded (2 percent) are less likely to say that they voted for Bündnis 90 / Die Grünen (Green Party) at the last general election than the digital addicts (11 percent), the late adopters (8 percent) or the traditional users (7 percent).

We conclude that the composition of our IT literacy classes differs with regard to the socio-demographic characteristics, political attitudes, and voting behavior of sample units. We thus find support for hypothesis 1.

To investigate whether classes of IT literacy – in addition to general predictors of nonresponse – correlate with sample units' propensity to respond to the first online panel wave of the GIP, we run several logistic regression models. The detailed regression results can be found in table 3.6. In additional figures, we visualize the results from table 3.6 as coefficient plots (see figure 3.1 and 3.2).

Table 3.6: Logistic regression on response to the first online panel wave.

	Model 1 general indicators		Model 2 +IT literacy		Model 3 +interactions	
	$\hat{\beta}$	Std. err.	$\hat{\beta}$	Std. err.	$\hat{\beta}$	Std. err.
Demographics						
Age	0.10***	(0.02)	0.08***	(0.02)	0.30***	(0.08)
Age^2	−0.00***	(0.00)	−0.00***	(0.00)	−0.00***	(0.00)
Being female	−0.10	(0.07)	0.01	(0.08)	0.03	(0.08)
Ref. Low educational level						
Medium educational level	0.51***	(0.09)	0.33***	(0.10)	0.31**	(0.10)
High educational level	1.06***	(0.10)	0.76***	(0.10)	0.75***	(0.10)
Currently pupil	1.58***	(0.38)	1.24**	(0.39)	1.13**	(0.39)
Ref. Single household						
Two hh members	0.15	(0.09)	−0.09	(0.09)	0.47	(0.29)
Three and more hh members	0.00	(0.09)	−0.26**	(0.10)	0.08	(0.57)
Ref. >500'inhabitants, core						
>500'inhabitants, periphery	0.03	(0.14)	0.02	(0.14)	0.01	(0.14)
100'– 500'inhabitants, core	0.21	(0.12)	0.22	(0.12)	0.23	(0.12)
100'– 500'inhabitants, periphery	0.07	(0.12)	0.13	(0.12)	0.13	(0.12)
50'– 100'inhabitants, core	−0.42	(0.25)	−0.26	(0.26)	−0.27	(0.27)
50'– 100'inhabitants, periphery	0.07	(0.14)	0.14	(0.14)	0.12	(0.14)
20'– 50'inhabitants, periphery	−0.09	(0.14)	−0.01	(0.14)	−0.00	(0.14)
5'– 20'inhabitants	−0.30*	(0.14)	−0.21	(0.15)	−0.22	(0.15)
2'– 5'inhabitants	−0.04	(0.21)	0.06	(0.23)	0.06	(0.23)
<2'inhabitants	0.19	(0.24)	0.31	(0.27)	0.31	(0.27)
Ref. Low political interest						
Some political interest	0.13	(0.13)	0.10	(0.13)	0.11	(0.13)
High political interest	0.34*	(0.13)	0.21	(0.14)	0.20	(0.14)
Ref. Voters						
Nonvoters	−0.56***	(0.11)	−0.54***	(0.11)	−0.56***	(0.11)
Not eligible	−1.01***	(0.15)	−0.95***	(0.16)	−0.98***	(0.15)
Don't know	−0.27	(0.15)	−0.30*	(0.15)	−0.28	(0.15)
Refused to answer	−0.62***	(0.12)	−0.61***	(0.12)	−0.61***	(0.12)
Ref. Digitally excluded						
Digital addicts			2.33***	(0.17)	7.50***	(2.21)
Late adopters			1.84***	(0.16)	6.93**	(2.24)
Traditional users			1.36***	(0.26)	8.35**	(2.83)
*Ref. Digitally excluded*age*						
Digital addicts*age					−0.26**	(0.08)
Late adopters*age					−0.23**	(0.08)
Traditional users*age					−0.32**	(0.12)
*Ref. Digitally excluded*age^2*						
Digital addicts*age^2					0.00***	(0.00)
Late adopters*age^2					0.00**	(0.00)
Traditional users*age^2					0.00**	(0.00)
*Ref. Digitally excluded*single household*						
Digital addicts*two hh members					−0.62	(0.32)
Digital addicts*three and more hh members					−0.30	(0.58)
Late adopters*two hh members					−0.55	(0.32)
Late adopters*three and more hh members					−0.46	(0.59)
Traditional users*two hh members					−1.49**	(0.57)
Traditional users*three and more hh members					−0.52	(0.78)
Base	3,970		3,970		3,970	
McFadden Pseudo R^2	0.08		0.13		0.13	

NOTE. – Design weights to account for unequal selection probabilities were applied. 9 cases deleted because they failed to respond to all questions used in the model.
$\hat{\beta}$ = coefficients, Std. err. = standard errors, hh = household, Ref. = Reference category
* p < 0.05, ** p < 0.01, *** p < 0.001.

The first model includes the general predictors of survey unit nonresponse age, gender, educational level, household size, urbanity, political interest, and electoral participation. We find that the propensity to participate in the first wave of the online panel significantly increases with age and educational level. Furthermore, the coefficient for a squared age term is negative and significant, suggesting an initial rise in the response propensity with increasing age and then a fall when reaching an older age. In addition, we find significant effects of political interest and voting behavior. Persons with high political interest are more likely to participate in the online panel than persons with some or low political interest. And voters are significantly more likely to respond than nonvoters.

In the second model, we extend the analysis by our classes of IT literacy in order to investigate whether IT literacy contributes to explaining online panel response in addition to the general predictors. By adding the classes of IT literacy the pseudo R^2 of the model increases from 0.08 to 0.13, indicating an improvement in the model fit (see table 3.6). The LR test on the two nested models shows that the predictor IT literacy significantly increases the fit of the model ($LR\,\chi^2 = 212.70$, $d.f. = 3$, $p = 0.00$). Furthermore, we find some indication for the model with the multi-dimensional classification of IT literacy fitting the data better than the one with the binary classification of being online and being offline ($AIC = 4,818$, $BIC = 4,987$ versus $AIC = 4,850$, $BIC = 5,008$; results not shown), though non-nested non-linear models cannot be compared conclusively. Therefore, IT literacy explains unit nonresponse in addition to general predictors, and we find support for our hypothesis 2.

Figure 3.1 presents a coefficient plot in which we further investigate the effect of the different classes of IT literacy on the response (based on table 3.6 model 2). We find that digital addicts, late adopters, and traditional users are significantly more likely to respond to the first wave of the online panel than the digitally excluded. The response propensity is highest for the digital addicts, followed by the late adopters, and the traditional users. Even though we see an overlap in confidence intervals for digital addicts, late adopters and traditional users, varying the reference categories (results not shown) reveals a significant difference between digital addicts and all other classes of IT literacy (for a discussion on overlapping confidence intervals and significant differences see Schenker and Gentleman, 2001). These results support our hypothesis that a person's propensity to respond to an online panel differs by the IT literacy class that they belong

to (hypothesis 3).

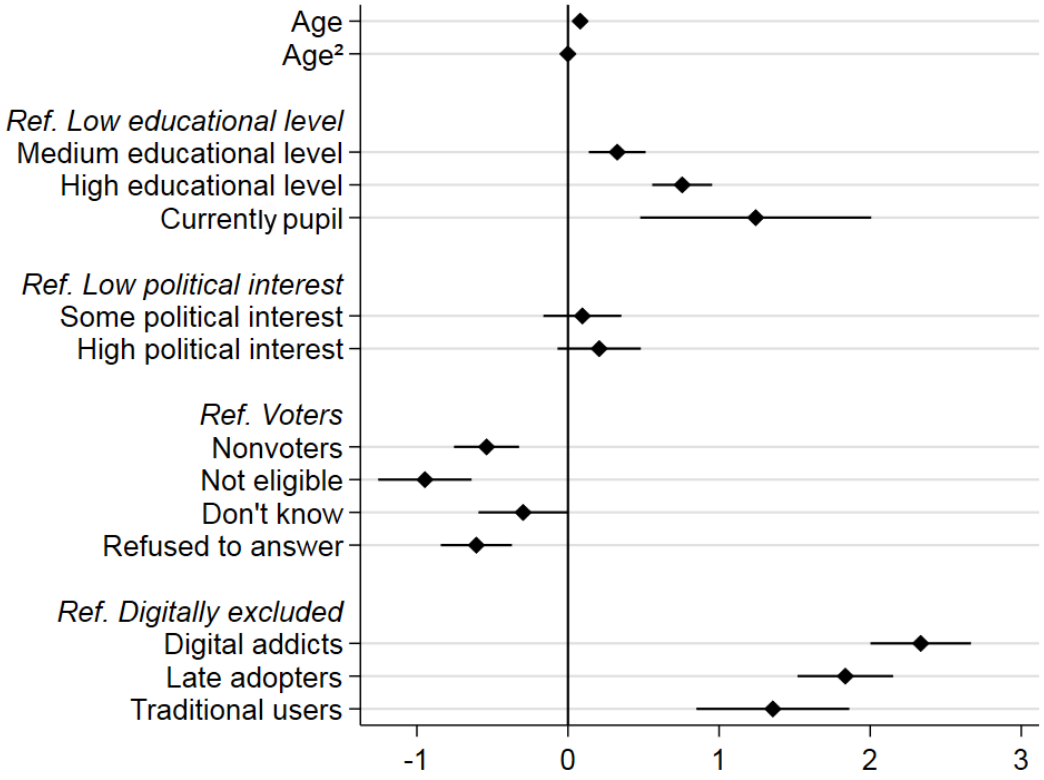


Figure 3.1: Logistic regression of sample units' characteristics on response to the first online panel wave. Based on model 2 in table 3.6. Plot markers are coefficients, and horizontal spikes are 95% confidence intervals.

To investigate further whether IT literacy adds to explaining the response to an online panel, we extend the second model by interaction terms between IT literacy and each of the significant general predictors of model 2 (see table 3.6 model 3). We find significant interactions of IT literacy with age and with household size. The interactions mean that IT literacy is not only directly associated with response to the online panel. In addition, the effect of IT literacy on nonresponse is differentially strong for people with different background characteristics. However, while the interaction between IT literacy and household size is only significant for one subgroup combination, the interaction of the IT literacy classes with age is consistent both in its direction and in its significance. It thus warrants further investigation.

The effect of age on response propensity clearly differs by class of IT literacy. The slopes of the predicted response probabilities differ across the four classes of IT literacy (see figure 3.2). For example, while the young digitally excluded sample units are very unlikely to participate in the online panel, the response propensity increases for the middle-aged but decreases again for the older digitally excluded. In contrast, the response propensity for digital addicts rises constantly with age. For traditional users and late adopters, the propensity to respond to the online panel is independent of age.

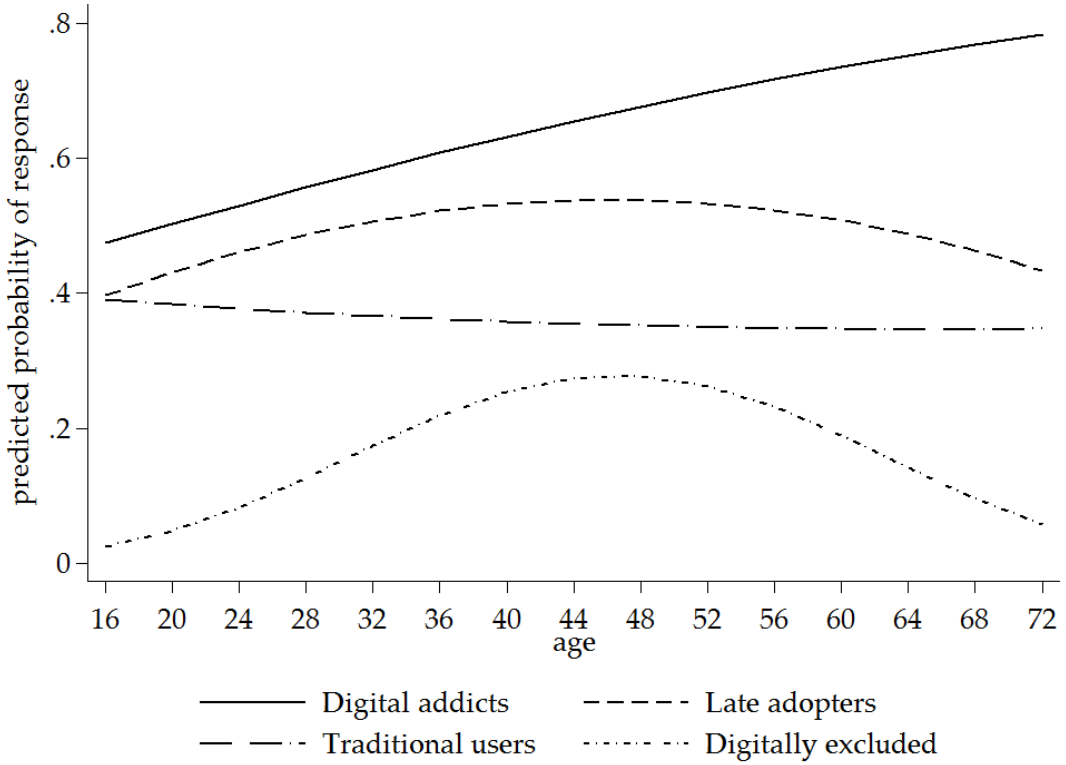


Figure 3.2: Predicted probabilities of response to the first online panel wave plotted against age, separately for IT literacy classes. Based on model 3 in table 3.6.

3.6.3 Online panel attrition by IT literacy

Our last set of analyses concerns the added value of IT literacy in explaining participation across online panel waves (i.e., attrition). We find a significantly negative effect of

wave and a significantly positive effect of wave squared, indicating that with ongoing panel waves the probability of initial panel members to respond to the wave requests becomes lower, but this decrease in response propensity flattens after about 1.5 years (wave 9, see table 3.7 model 1).

Table 3.7: Logistic regression on participation across panel waves.

	Model 1 retention		Model 2 +interactions	
	$\hat{\beta}$	Std. err.	$\hat{\beta}$	Std. err.
Panel waves				
Wave	−0.33***	(0.01)	−0.15**	(0.07)
Wave^2	0.01***	(0.00)	0.00	(0.00)
Ref. Digitally excluded				
Digital addicts	0.21	(0.23)	1.06***	(0.37)
Late adopters	−0.01	(0.23)	0.91**	(0.37)
Traditional users	−0.09	(0.34)	0.57	(0.57)
*Ref. Digitally excluded*wave*				
Digital addicts*wave			−0.19***	(0.07)
Late adopters*wave			−0.19***	(0.07)
Traditional users*wave			−0.18	(0.12)
*Ref. Digitally excluded*wave^2*				
Digital addicts*wave^2			0.01**	(0.00)
Late adopters*wave^2			0.01**	(0.00)
Traditional users*wave^2			0.01	(0.01)
Base	26,832		26,832	
McFadden Pseudo R^2	0.03		0.03	

NOTE. – Wave 3 was left out of the analysis to account for the late recruitment of offliners.
$\hat{\beta}$ = coefficients, Std. err. = clustered standard errors, Ref. = Reference category
* p < 0.05, ** p < 0.01, *** p < 0.001

To investigate separate time-trends for each class of IT literacy, we include interaction terms of waves and IT literacy in our model (table 3.7 model 2). We see a significant difference between the probability of participation across panel waves in the classes of the digital excluded and the digital addicts as well as the digital excluded and the late adopters. These results are also robust when we estimate a random effects model with clustered sandwich estimators for general errors instead of the logistic regressions

presented here.

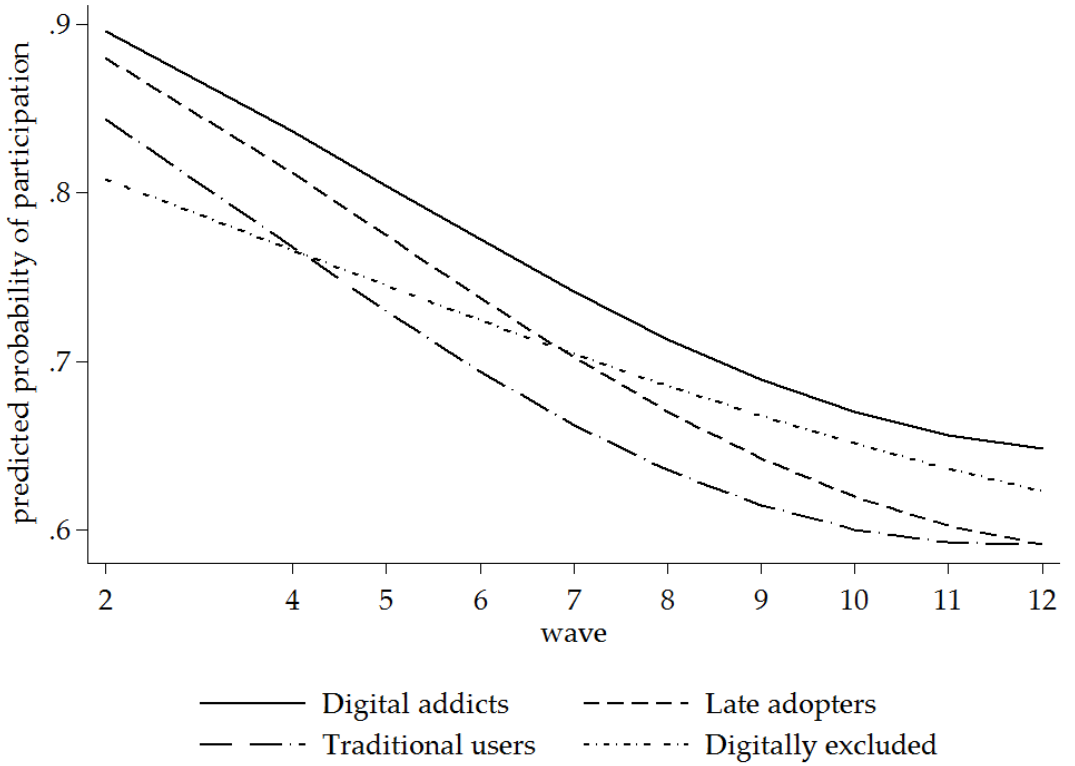


Figure 3.3: Predicted probabilities of participation in the online panel plotted against panel waves, separately for IT literacy. Based on model 2 in table 3.7. Wave 3 is purposively missing from the analysis to account for the late recruitment of offliners.

The interaction terms are visualized in figure 3.3, where we plotted predicted probabilities of attrition against panel waves, separately for each IT literacy class. Figure 3.3 shows that the classes of IT literacy differ in their probability of participate in the online waves over time. The attrition curve for the digital addicts is higher than the curve for any other class of IT literacy. Furthermore, while for the digital addicts the probability of participating in a panel wave decreases relatively steeply at the beginning, for late adopters and traditional users it stabilizes after about ten waves. In contrast, the functional form of the attrition curve for the digitally excluded is almost linear with little change over time. The slope of the attrition curve flattens more quickly for the late adopters and traditional users than for digital addicts and digitally excluded. As a consequence, after

wave 4, the digitally excluded are more likely to participate in a panel wave than the traditional users. After wave 7, the digitally excluded are also more likely to participate in a panel wave than the late adopters. These results are in line with our hypothesis that attrition rates differ by class of IT literacy (hypothesis 4).

3.7 Discussion

Research into the selectivities of probability-based online panels has thus far considered the digital divide to be a binary phenomenon: persons were either offline or online. In this article, we draw attention to a change in the digital divide as a result of increasing internet penetration rates and an increasing diversification in people's participation in digital developments. We propose a more fine-grained classification of IT literacy to investigate unit nonresponse and unit nonresponse bias in probability-based online panels. Our analyses are based on data from the 2014 face-to-face recruitment and subsequent online interviews of the German Internet Panel (GIP), a probability-based online panel of the general population.

In a first step, we extend the binary differentiation of having internet access versus not having internet access into a multi-dimensional classification covering aspects of the digital access divide, the digital device divide, the digital usage divide, as well as attitudes towards technical innovations. A latent class analysis identifies four classes of IT literacy: digital addicts, late adopters, traditional users, and digitally excluded. Digital addicts have a computer and internet at home, own new technological devices, such as smartphones, use the internet on a daily basis, and consider technology important. Late adopters have a computer and internet at home, and some of them own a smartphone. Many late adopters use the internet on a daily basis, but they attach little importance to technology. Traditional users have computers without internet or no computers at all at home, do not own a smartphone and consider technology unimportant. Nonetheless, most traditional users use the internet on a daily basis, presumably at work. Finally, the digitally excluded never use the internet, do not own a smartphone, and attach little or no importance to technology.

Investigating the characteristics of the four classes of IT literacy, their probability to become online panel members, and their probability to continue participation in online

panel waves, we find support for each of our four hypotheses.

First, we show that sample units that belong to different IT literacy classes also significantly differ in key survey variables, in particular in their socio-demographic characteristics, political attitudes, and voting behavior.

Second, we show that our multi-dimensional conceptualization of IT literacy, alongside the general predictors of unit nonresponse, explains panel membership in the GIP. An increase in the pseudo R^2 when adding classes of IT literacy to the model indicates an added value in the classes of IT literacy for predicting participation in the first online panel wave.

Third, we find that sample units' propensity to participate in the first online panel wave varies by their class of IT literacy. We find evidence that not only offliners but also certain groups among the online population are underrepresented in online panels. More generally, different IT literacy classes have different propensities to participate in the GIP: Digital addicts have a higher probability to respond to the online panel than late adopters, traditional users, and the digitally excluded. In addition, IT literacy interacts with age. The functional form and slopes of the association between age and response differ across the four classes of IT literacy. For example, while we observe an inversely U-shaped relationship for the digitally excluded, the response propensity for digital addicts rises constantly with age. In contrast, for traditional users and late adopters response to the online panel appears not to be related to age.

Fourth, we find evidence that different IT literacy classes are associated with different attrition patterns. The functional forms of the estimated logistic regression curves differ by class of IT literacy: While the digitally excluded are hesitant starters, after a few waves, they are more likely to participate every two months in online interviews than late adopters and traditional users. These findings suggest that the digitally excluded are very committed to the online panel after having been equipped with the necessary devices. This commitment may stem from a feeling of obligation resulting from the investments made by providing them with computer and internet access.

3.8 Conclusions

With the ubiquity of online surveys in social, political, and economic research, understanding the underlying selection mechanisms is of key interest to both methodologists and researchers using survey data. Recent years have seen the establishment of probability-based online panels, which are representative of the general population, many of which are accompanied with research programs into the coverage and nonresponse biases that may stem from the online mode (see for example Blom et al., 2017; de Vos, 2010; Leenheer and Scherpenzeel, 2013; Revilla et al., 2016). Most of this research, however, investigate whether it "pays off" to integrate persons who have no computer and/or internet (so-called offliners) in the sample to reduce coverage bias; it overlooks potential biases that arise from groups within the online and offline population that may be differentially likely to participate in the online panel due to unit nonresponse.

The results presented in this paper identify four different classes of IT literacy among the general population in Germany. These four classes differ in their personal characteristics, the initial response to the online panel, and longitudinal attrition patterns. Following Groves' (2006) common cause model, we detect nonresponse bias with respect to IT literacy in the GIP data. However, this finding is accompanied by good news: We can use our IT literacy indicator to reduce the detected bias. In addition, the knowledge gained from our research can inform fieldwork monitoring when recruiting and maintaining probability-based online panels. Well-designed fieldwork monitoring procedures (for example in responsive designs see Groves and Heeringa, 2006) may even overcome this nonresponse bias in the future.

While we were very fortunate to receive access to the detailed recruitment data of the GIP, our results are also limited by the data available. The GIP face-to-face recruitment interview was neither designed to measure IT literacy nor envisaged to facilitate the research we conducted. We hope that our results will inform the design of future recruitment interviews of the GIP and other probability-based online panels. A purposively developed recruitment questionnaire would allow to better capture the different aspects of IT literacy. In particular, it may enable a further sub-division of the large and diverse class of digital addicts. In 2014 this class already contained the majority of GIP sample; increasing digitalization is likely to further strengthen and diversify this class, a

process we should capture and analyze. Finally, more detailed data on IT literacy during the recruitment process may improve fieldwork monitoring processes.

More generally, our research contains key messages for both survey practitioners and analysts. To survey practitioners setting up probability-based online panels, we recommend monitoring participation in the online panel across classes of IT literacy in the course of the recruitment and panel maintenance processes. Covering IT literacy classes during interviewer training or considering IT literacy classes in the amount and timing of reminders might be an effective strategy to decrease differential unit nonresponse in the first place. For example, using more experienced interviewers for digitally excluded who are hard to recruit but attrite less, or sending more reminders to late adopters after six panel waves might reduce unit nonresponse. For analysts, classes of IT literacy may be valuable when developing effective weights, because they comprise information that reaches beyond general predictors of unit nonresponse and is correlated with key survey outcomes. Future research should look into the development of such nonresponse weights.

On a final note, online surveys seem to be here to stay. Their data facilitate research on a diversity of essential research topics, in the case of the GIP, for example, the political economy of reforms. Understanding the mechanism leading to selectivities in the data of online surveys and discovering ways to adjust for such selectivities will enable a better understanding of political, social, and economic processes and is a prerequisite for making substantive research count.

Acknowledgments

This work was supported by the German Research Foundation (DFG, [SFB 884]). The authors gratefully acknowledge support from the Collaborative Research Center (SFB) 884 "Political Economy of Reforms" at the University of Mannheim. The German Internet Panel is the central data collection of the SFB 884 (project Z1). We especially thank Vlad Achimescu and Florian Keusch for comments on an earlier version of the paper. The authors would further like to thank Daniela Ackermann-Piek, Julian Axenfeld, Christian Bruch, Carina Cornesse, Barbara Felderer, Franziska Gebhard, Susanne Helmschrott, Tobias Rettig and Ulrich Krieger.

References

Antoun, C. (2015). Who are the Internet Users, Mobile Internet Users, and Mobile-Mostly Internet Users?: Demographic Differences across Internet-Use Subgroups in the U. S. In Toninelli, D., Pinter, R., and de Pedraza, P., editors, *Mobile Research Methods: Opportunities and Challenges of Mobile Research Methodologies*, chapter 7, pages 99–117. Ubiquity Press, London, UK.

Baker, R., Blumberg, S. J., Brick, J. M., Couper, M. P., Courtright, M., Dennis, J. M., Dillman, D., Frankel, M. R., Garland, P., Groves, R. M., Kennedy, C., Krosnick, J., Lavrakas, P. J., Lee, S., Link, M., Piekarski, L., Rao, K., Thomas, R. K., and Zahs, D. (2010). Research Synthesis: AAPOR Report on Online Panels. *Public Opinion Quarterly*, 74(4):711–781.

Barron, M. and Dennis, J. M. (2016). *Multi-Client Household Panel Quality – The Case of AmeriSpeak*. Presentation at the 2016 American Association of Public Opinion Research, Austin, TX, USA.

Best, S. J., Krueger, B., Hubbard, C., and Smith, A. (2001). An Assessment of the Generalizability of Internet Surveys. *Social Science Computer Review*, 19(2):131–145.

Bethlehem, J. (2010). Selection Bias in Web Surveys. *International Statistical Review*, 78(2):161–188.

Bethlehem, J. and Stoop, I. (2007). Online Panels – a Paradigm Theft. In Trotman, M., Burrell, T., Andertson, K., Basi, G., Cooper, M. P., Morris, K., Birks, D., Johnson, A. J., Baker, R., Rigg, M., Taylor, S., and Westlake, A., editors, *The Challenges of a Changing World*, pages 113–131, Southampton, UK. Association for Survey Computing.

Blank, G. (2016). The Digital Divide among Twitter Users and Its Implications for Social Research. *Social Science Computer Review*, 35(6):679–697.

Blank, G. and Groselj, D. (2014). Dimensions of Internet Use: Amount, Variety, and Types. *Information, Communication & Society*, 17(4):417–435.

Blom, A. G., Bosnjak, M., Cornilleau, A., Cousteaux, A.-S., Das, M., Douhou, S., and Krieger, U. (2016). A Comparison of Four Probability-Based Online and Mixed-Mode Panels in Europe. *Social Science Computer Review*, 34(1):8–25.

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., Herzing, J. M. E., Cornesse, C., Sakshaug, J. W., Krieger, U., and Bossert, D. (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, M., Dannwolf, T., Enderle, T., Schaurer, I., Struminskaya, B., Tanner, A., and 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*. https://doi.org/10.1177/0894439317697949.

Bosnjak, M., Haas, I., Galesic, M., Kaczmirek, L., Bandilla, W., and Couper, M. P. (2013). Sample Composition Discrepancies in Different Stages of a Probability-Based Online Panel. *Field Methods*, 25(4):339–360.

Brandtzæg, P. B. (2010). Towards a Unified Media-User Typology (MUT): A Meta-Analysis and Review of the Research Literature on Media-User Typologies. *Computers in Human Behavior*, 26(5):940–956.

Brandtzæg, P. B., Heim, J., and Karahasanović, A. (2011). Understanding the New Digital Divide – A Typology of Internet Users in Europe. *International Journal of Human-Computer Studies*, 69(3):123–138.

Brick, J. M. and Tourangeau, R. (2017). Responsive Survey Designs for Reducing Nonresponse Bias. *Journal of Official Statistics*, 33(3):735–752.

Couper, M. P., Kapteyn, A., Schonlau, M., and Winter, J. (2007). Noncoverage and Nonresponse in an Internet Survey. *Social Science Research*, 36(1):131–148.

de Bruijne, M. and Wijnant, A. (2014). Mobile Response in Web Panels. *Social Science Computer Review*, 32(6):728–742.

de Vos, K. (2010). Representativeness of the LISS-Panel 2008, 2009, 2010.

Dever, J. A., Rafferty, A., and Valliant, R. (2008). Internet Surveys: Can Statistical Adjustments Eliminate Coverage Bias? *Survey Research Methods*, 2(2):47–60.

DiMaggio, P., Hargittai, E., Celeste, C., and Shafer, S. (2004). From Unequal Access to Differentiated Use: A Literature Review and Agenda for Research on Digital Inequality. *Social Inequality*, pages 355–400.

Dutton, W. H. and Blank, G. (2014). The Emergence of Next Generation Internet Users. *International Economics and Economic Policy*, 11(1-2):29–47.

European Commission (2014). Special Eurobarometer 414, Wave EB81.1, E-Communications and Telecom Single Market Household Survey. http://ec.europa.eu/COMMFrontOffice/publicopinion/index.cfm/ResultDoc/download/DocumentKy/57810.

European Commission (2016). Special Eurobarometer 438, Wave EB84.2, E-Communications and Telecom Single Market Household Survey. http://ec.europa.eu/COMMFrontOffice/publicopinion/index.cfm/ResultDoc/download/DocumentKy/72564.

Friemel, T. N. (2016). The Digital Divide Has Grown Old: Determinants of a Digital Divide among Seniors. *New Media & Society*, 18(2):313–331.

Fuchs, M. and Busse, B. (2009). The Coverage Bias of Mobile Web Surveys across European Countries. *International Journal of Internet Science*, 4(1):21–33.

Gelman, A., Goel, S., Rothschild, D., and Wang, W. (2016). High-Frequency Polling with Non-Representative Data. In Schill, D., Kirk, R., and Jasperson, A. E., editors, *Political Communication in Real Time: Theoretical and Applied Research Approaches*, chapter 5, pages 89–105. Routledge, New York, NY.

Gould, W. W. (1995). sg34: Jackknife Estimation. *Stata Technical Bulletin*, 24:25–29.

Greenlaw, C. and Brown-Welty, S. (2009). A Comparison of Web-Based and Paper-Based Survey Methods Testing Assumptions of Survey Mode and Response Cost. *Evaluation Review*, 33(5):464–480.

Groves, R. M. (2006). Nonresponse Rates and Nonresponse Bias in Household Surveys. *Public Opinion Quarterly*, 70(5):646–675.

Groves, R. M. and 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.

Groves, R. M., Presser, S., and Dipko, S. (2004). The Role of Topic Interest in Survey Participation Decisions. *Public Opinion Quarterly*, 68(1):2–31.

Guo, R. X., Dobson, T., and Petrina, S. (2008). Digital Natives, Digital Immigrants: An Analysis of Age and ICT Competency in Teacher Education. *Journal of Educational Computing Research*, 38(3):235–254.

Hardigan, P. C., Succar, C. T., and Fleisher, J. M. (2012). An Analysis of Response Rate and Economic Costs between Mail and Web-Based Surveys among Practicing Dentists: A Randomized Trial. *Journal of Community Health*, 37(2):383–394.

Holmes, J. (2011). Cyberkids or Divided Generations? Characterising Young People's Internet Use in the UK with Generic, Continuum or Typological Models. *New Media & Society*, 13(7):1104–1122.

Hong, S., Thong, J. Y. L., and Tam, K. Y. (2006). Understanding Continued Information Technology Usage Behavior: A Comparison of Three Models in the Context of Mobile Internet. *Decision Support Systems*, 42(3):1819–1834.

Hoogendoorn, A. and Daalmans, J. (2009). Nonresponse in the Recruitment of an Internet Panel Based on Probability Sampling. *Survey Research Methods*, 3(2):59–72.

Kaplowitz, M. D., Hadlock, T. D., and Levine, R. (2004). A Comparison of Web and Mail Survey Response Rates. *Public Opinion Quarterly*, 68(1):94–101.

Knoef, M. and de Vos, K. (2009). *The Representativeness of LISS, an Online Probability Panel*. CentERdata, Tilburg, NL. https://www.researchgate.net/profile/Marike_Knoef/publication/242742051_The_representativeness_of_LISS_an_online_probability_panel/links/0f3175339ae828f081000000.pdf.

Kreuter, F., Olson, K., Wagner, J., Yan, T., Ezzati-Rice, T. M., Casas-Cordero, C., Lemay, M., Peytchev, A., Groves, R. M., and 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.

Kreuter, F., Presser, S., and Tourangeau, R. (2008). Social Desirability Bias in CATI, IVR, and Web Surveys: The Effects of Mode and Question Sensitivity. *Public Opinion Quarterly*, 72(5):847–865.

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

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.

Little, R. J. A. (1988). Missing-Data Adjustments in Large Surveys. *Journal of Business & Economic Statistics*, 6(3):287–296.

Lugtig, P. (2014). Panel Attrition: Separating Stayers, Fast Attriters, Gradual Attriters, and Lurkers. *Sociological Methods & Research*, 43(4):699–723.

McCutcheon, A. L. (2002). Basic Concepts and Procedures in Single- and Multiple-Group Latent Class Analysis. In Hagenaars, J. A. and McCutcheon, A. L., editors, *Applied Latent Class Analysis*, chapter 2, pages 54–88. Cambridge University Press, Cambridge, UK.

Millar, M. M. and Dillman, D. A. (2011). Improving Response to Web and Mixed-Mode Surveys. *Public Opinion Quarterly*, 75(2):249–269.

Mohorko, A., de Leeuw, E. D., and Hox, J. J. (2013). Internet Coverage and Coverage Bias in Europe: Developments across Countries and over Time. *Journal of Official Statistics*, 29(4):609–622.

Morris, T. P., White, I. R., and Royston, P. (2014). Tuning Multiple Imputation by Predictive Mean Matching and Local Residual Draws. *BMC Medical Research Methodology*, 14(75):1–13.

Nylund, K. L., Asparouhov, T., and Muthén, B. O. (2007). Deciding on the Number of Classes in Latent Class Analysis and Growth Mixture Modeling: A Monte Carlo Simulation Study. *Structural Equation Modeling*, 14(4):535–569.

Olson, K., Smyth, J. D., and Wood, H. M. (2012). Does Giving People Their Preferred Survey Mode Actually Increase Survey Participation Rates? An Experimental Examination. *Public Opinion Quarterly*, 76(4):611–635.

Payne, J. and Barnfather, N. (2012). Online Data Collection in Developing Nations: An Investigation into Sample Bias in a Sample of South African University Students. *Social Science Computer Review*, 30(3):389–397.

Peytchev, A., Riley, S., Rosen, J., Murphy, J., and Lindblad, M. (2010). Reduction of Nonresponse Bias Through Case Prioritization. *Survey Research Methods*, 4(1):21–29.

Pforr, K. and Dannwolf, T. (2017). What Do We Lose With Online-only Surveys? Estimating the Bias in Selected Political Variables Due to Online Mode Restriction. *Statistics, Politics and Policy*, 8(1):105–120.

Porter, C. E. and Donthu, N. (2006). Using the Technology Acceptance Model to Explain How Attitudes Determine Internet Usage: The Role of Perceived Access Barriers and Demographics. *Journal of Business Research*, 59(9):999–1007.

Quenouille, M. H. (1956). Notes on Bias in Estimation. *Biometrika*, 43(3/4):353–360.

Revilla, M., Cornilleau, A., Cousteaux, A.-S., Legleye, S., and de 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.

Rice, R. E. and Katz, J. E. (2003). Comparing Internet and Mobile Phone Usage: Digital Divides of Usage, Adoption, and Dropouts. *Telecommunications Policy*, 27(8):597–623.

Robinson, J. P., Neustadtl, A., and Kestnbaum, M. (2002). The Online 'Diversity Divide': Public Opinion Differences among Internet Users and Nonusers. *IT & Society*, 1(1):284–302.

Rookey, B. D., Hanway, S., and Dillman, D. A. (2008). Does a Probability-Based Household Panel Benefit from Assignment to Postal Response as an Alternative to Internet-Only? *Public Opinion Quarterly*, 72(5):962–984.

Schenker, N. and Gentleman, J. F. (2001). On Judging the Significance of Differences by Examining the Overlap between Confidence Intervals. *The American Statistician*, 55(3):182–186.

Schnell, R. (1997). *Nonresponse in Bevölkerungsumfragen: Ausmaß, Entwicklung und Ursachen*. Leske + Budrich, Opladen, DE.

Schnell, R., Noack, M., and Torregoza, S. (2017). Differences in General Health of Internet Users and Non-Users and Implications for the Use of Web Surveys. *Survey Research Methods*, 11(2):103–123.

Schonlau, M., Fricker, R. D., and Elliott, M. N. (2002). *Conducting Research Surveys via E-Mail and the Web*. Rand Coperation, Santa Monica, CA.

Shih, T.-H. and Fan, X. (2007). Response Rates and Mode Preferences in Web-Mail Mixed-Mode Surveys: A Meta-Analysis. *International Journal of Internet Science*, 2(1):59–82.

Singer, E. (2011). Toward a Benefit-Cost Theory of Survey Participation: Evidence, Further Tests, and Implications. *Journal of Official Statistics*, 27(2):379–392.

Slegers, K., van Boxtel, M. P. J., and Jolles, J. (2012). Computer Use in Older Adults: Determinants and the Relationship with Cognitive Change over a 6 Year Episode. *Computers in Human Behavior*, 28(1):1–10.

Stempel, I. G. H., Hargrove, T., and Bernt, J. P. (2000). Relation of Growth of Use of the Internet to Changes in Media Use from 1995 to 1999. *Journalism & Mass Communication Quarterly*, 77(1):71–79.

Sterrett, D., Malato, D., Benz, J., Tompson, T., and English, N. (2017). Assessing Changes in Coverage Bias of Web Surveys in the United States. *Public Opinion Quarterly*, 81(S1):338–356.

van Deursen, A. J. A. M. and van Dijk, J. A. G. M. (2014). The Digital Divide Shifts to Differences in Usage. *New Media & Society*, 16(3):507–526.

van Deursen, A. J. A. M. and van Dijk, J. A. G. M. (2015). Internet Skill Levels Increase, but Gaps Widen: A Longitudinal Cross-Sectional Analysis (2010–2013) among the Dutch Population. *Information, Communication & Society*, 18(7):782–797.

Vermunt, J. K. and Magidson, J. (2016). *Technical Guide for Latent GOLD 5.1: Basic, Advanced, and Syntax.* Statistical Innovations Inc., Belmont, MA.

Wang, Q. E., Myers, M. D., and Sundaram, D. (2013). Digital Natives and Digital Immigrants: Towards a Model of Digital Fluency. *Business & Information Systems Engineering*, 5(6):409–419.

Wei, L. and Hindman, D. B. (2011). Does the Digital Divide Matter More? Comparing the Effects of New Media and Old Media Use on the Education-Based Knowledge Gap. *Mass Communication and Society*, 14(2):216–235.

West, B. T. and Blom, A. G. (2016). Explaining Interviewer Effects: A Research Synthesis. *Journal of Survey Statistics and Methodology*, 5(2):175–211.

Yeager, D. S., Krosnick, J. A., Chang, L., Javitz, 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.

Zhang, C., Callegaro, M., and Thomas, M. (2008). More than the Digital Divide? Investigating the Differences between Internet and Non–Internet Users on Attitudes and Behaviors. Paper presented at Midwest Association for Public Opinion Research (MAPOR) 2008 Conference, Chicago, IL.

Zillmann, D., Schmitz, A., Skopek, J., and Blossfeld, H.-P. (2014). Survey Topic and Unit Nonresponse. *Quality & Quantity*, 48(4):2069–2088.

Chapter 4

Paper III

Investigating alternative interface designs for long-list questions

Abstract

This study extends previous research on visual design effects of long-list questions. I evaluate three different response formats for long-list questions with regard to differences in response burden and data quality. At the example of the highest educational qualification, I compare a long list of radio buttons, a combo box (combination of a text box and a drop-down box), and a search tree (nested list of response options). For this purpose, I conducted a split-ballot experiment with three conditions in the Innovation Sample of the German Socio-Economic Panel (SOEP-IS 2014). The findings indicate that combo boxes are most effective at reducing response burden, while search trees and long lists produce higher data quality in terms of post-coding effort. Consequently, there is a trade-off between a reduction of response burden and an increase of post-survey coding effort across the three interface designs.

Investigating alternative interface designs for long-list questions

4.1 Introduction

Questions with a large number of response options, such as questions on occupation, prescription drugs, country of residence, brands of product groups, field of study, or educational qualifications are difficult for respondents to answer and for survey practitioners to design (see Couper and Zhang, 2016; Keusch, 2014; Schierholz et al., 2017; Stern, 2008; Tijdens, 2015). Such questions are typically either asked as open-ended questions or as long-list questions. On the one hand, open-ended questions are challenging for survey organizations, because the collection of codeable data needs to be ensured (for visual design effects for open-ended input fields see Couper et al., 2011; Keusch, 2014; for the example of occupation see Conrad et al., 2016). On the other hand, open-ended questions can be less burdensome for respondents in terms of time to respond (see Stern, 2008). In contrast to open-ended questions, closed-ended question formats that contain long lists of answer options face the problem of the incompleteness of response options (see Fowler, 1995; Schneider, 2008), as the number of answer options is limited to page/screen space. In addition, the longer the list of answer options the more respondents have to read and the more likely it is that respondents are confused by answer options they might not know (see Lenzner et al., 2010). Thus, long lists of answer options can increase respondents' burden (see Fowler, 1995). Furthermore, closed-ended questions are likely to introduce response order effects (Krosnick and Alwin, 1987). Therefore, both open-ended questions and long-list questions challenge respondents and survey researchers in different ways.

Survey practitioners attempt to optimize long-list questions to facilitate respondents in their answer process by developing alternative interface designs. As an alternative solution to the established interface designs of long-list questions (e.g., standard text fields or lists of response options with radio buttons), new interface designs have emerged in computer-assisted surveys (for examples see Couper and Zhang, 2016; Funke and Reips, 2007; Tijdens, 2015). A general concern of introducing new interface designs compared to established interface designs is that respondents' answers are likely

to be different (see de Leeuw, 2005; Dillman et al., 2009; Tourangeau et al., 2010), as the visual presentation of response options affects response behavior (for examples see Christian and Dillman, 2004; Couper et al., 2004; Smyth et al., 2006; Redline and Dillman, 2002). Therefore, it seems prudent to evaluate alternative interface designs before they are implemented in surveys.

This study seeks to investigate the effect of two interface designs on response burden and data quality. For this purpose, the example of educational qualifications is used, as the interface design of this key survey variable is challenging survey organizations recently. Thus, this paper examines whether it is efficient to change the interface design of questions on educational qualifications from a long list of radio-buttons to a combo box or search tree.

4.2 Background

With technical advancements, alternative interface designs for long-list questions emerged. The first technological advancements allowed the grouping of response options by umbrella terms to support respondents' navigation through the offered response categories. In this context, the research examined respondents' efficiency when response categories were grouped. For instance, Stern (2008) and Redline et al. (2009) observe that conceptual grouping of response options (grouping by generic terms) results in longer response times and a higher amount of response editing compared to open-ended questions or long lists of answer options. In addition, other research has shown that long lists with radio buttons, drop-down boxes or type-ahead lookups face the problem of response order effects (see Couper, 2008; Couper et al., 2004; Healey, 2007; Stern, 2008, pp. 120-122).

The possibility to deploy plugin based programming languages (e.g., Java or HTML scripts) in survey software enabled various visual designs for long-list questions. For example, auto-completion text boxes (completes the entered text automatically) and auto-suggest text boxes (response suggestions are shown below the text field) emerged. However, research by Funke and Reips (2007) on these interface designs was inconclusive in terms of response times, item missings and dropouts.

With an increase of computer power search trees (nested list of response options)

and semantic matching tools (identifies semantic similarities in two structures, e.g., between a response and an entry in an underlying database) were developed to avoid multi-page filtering. Based on a non-experimental comparison of search trees with a semantic matching tool, Tijdens (2015, p. 10) recommends semantic matching tools over search trees when large numbers of response categories are necessary otherwise a search tree would fulfill the requirements. More recently, Couper and Zhang (2016) compared a standard text field, a drop-down box and a combo box (a combination of a text field and a drop-down box, also called combination box or database lookup) with each other. They conclude that the choice for one or the other interface design depends on the objectives of the questions, as both combo boxes and drop-down boxes reduce post-coding efforts, whereas standard text fields reduce response times (Couper and Zhang, 2016).

In the following, I focus on the interface design of the question on educational qualification, as the measurement of this key survey variable is challenging survey research. Educational qualification is a commonly used socio-economic variable in substantive research and hence, it is captured in almost all surveys of individuals (Smith, 1995). Thus, the measurement quality of an educational qualification is relevant for most researchers. Furthermore, there is an increase in complexity in measuring education in Germany and Europe due to rising differentiation of educational systems, such as national reforms following the Bologna Reform in Europe (Schneider, 2008). Yet, the closed-ended question format makes it challenging to present complex education systems (Schneider, 2008, p. 311), resulting in difficulties in recording the answers and in coding the answers into classification schemes of education (e.g., International Standard Classification of Education, ISCED). Therefore, closed-ended questions on educational qualifications become more difficult to design.

In summary, combo boxes and search trees are already used in computer-assisted surveys and yet, we do not know whether one or the other design format of long-list questions facilitates respondents in answering questions and how measurement is affected by these design formats. Therefore, this paper adds to our understanding of how to reduce response burden and increase measurement quality in self-administered surveys for long-list questions. In addition, the measurement of educational qualifications got more challenging. Hence, my study investigates whether there is an alternative to measuring educational qualifications which might improve the measurement of this key survey variable.

Using the example of educational qualification in Germany, I incorporated a radio-button list, a combo box, and a search tree in a computer-assisted panel survey. For this purpose, a split-ballot experiment with three conditions was conducted in the Innovation Sample of the German Socio-Economic Panel 2014 (SOEP-IS). First, I scrutinize whether presenting a long list of response options, a combo box or a search tree facilitates respondents in answering long-list questions. Second, I examine whether a long list of response options, a combo box or a search tree differ in data quality. Third, I examine whether differences across the experimental conditions emerge by respondent characteristics.

4.3 Alternative interface designs

To examine how we can improve the measurement of educational qualifications and help respondents in giving answers when faced with long-list questions, I compare three different interface designs[1] for the question on educational qualification[2]:

1. A long list of radio buttons with 28 response options out of which respondents choose an appropriate response (see figure 4.1). The educational qualifications were hierarchically ordered and categories for "other qualification" and "refuse to answer" were offered.

2. A 2-level search tree with 35 response options (see figure 4.2). Respondents see a single table with generic response categories. These response categories are ordered hierarchically. After the respondent has selected a generic response category, the response categories are displayed in a second table with more detailed response options (nested response options). If no selection was made in the search tree, respondents had to give an answer in a standard text field.

3. A combo box with 400 response options[3] (for illustration see figure 4.3). The combo box was the front end of a database query which was incorporated in the

[1]The combo box, and the search tree are based on Windows Presentation Foundation (WPF) technology and was recalled by the TNS infratest/Kantar survey software nipo. For further information see www.surveycodings.org/education.

[2]I present the German response options because translating German educational qualifications, is error-prone (Schneider et al., 2016b).

[3]The large number of response options is a result of including generic and specific terms for educa-

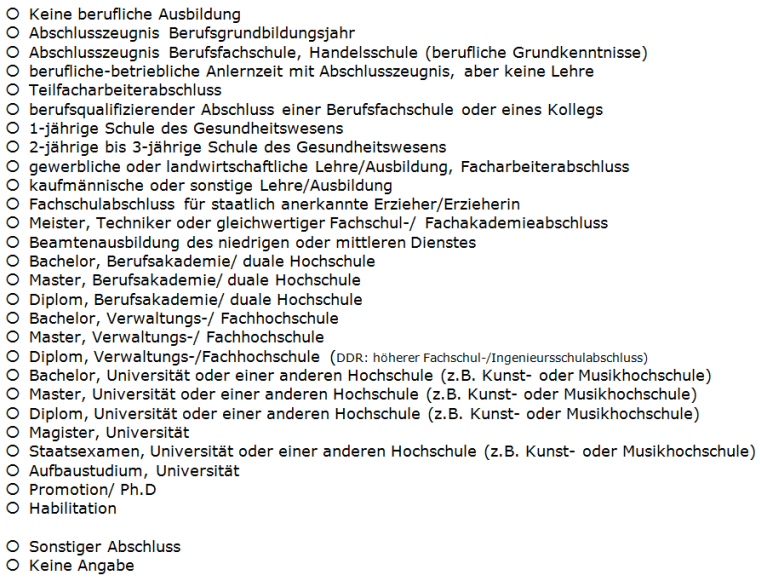


Figure 4.1: Established interface design - Long list with 28 response options.

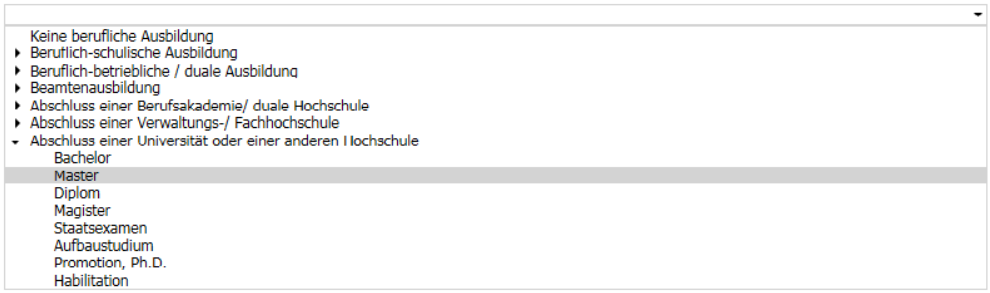


Figure 4.2: Alternative interface design - Search tree with 38 response options.

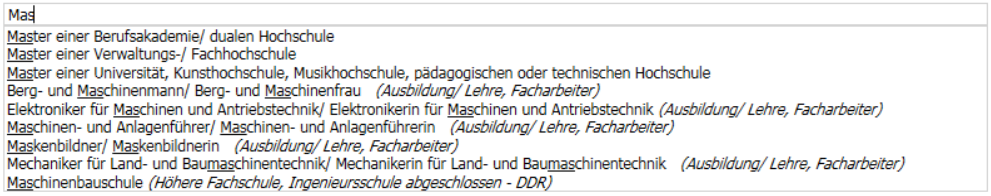


Figure 4.3: Alternative interface design - Combo box with over 400 response options.

survey procedure. Initially, the combo box does not differ from a standard text field. However, the moment the respondent starts typing, multiple suggestions for the most probable matches were offered below the text field (type suggestion). Each additional letter reduced the number of suggestions[4]. The respondents either typed in the response, selected a response from the list, or used a mix of both approaches.

When comparing the response options of the combo box or the search tree with the established question design – the long list with 28 response options asked in wave 2 and the list with 10 response options asked in wave 1 (for an illustration see figure A4.1) - we see that the combo box and the search tree allow for far more response options on a single page/screen than the lists of response options of the established question design.

The combo box combines the advantages of open-ended and closed-ended question formats and hence, overcomes the disadvantages of both question types (Funke and Reips, 2007). Initially, the combo box does not give respondents any guidance regarding the answer format. Thus, respondents are not restricted to the displayed response options and hence, respondents can give their intended response in open-ended questions (Fowler, 1995, p. 57). Therefore, I expect that the combo box facilitates respondents better in giving answers than the long list in terms of response burden.

Search trees do not give respondents as much flexibility in their response options as combo boxes do. By using search trees, one has to assume that respondents find a suitable response option in the offered responses. The concern with search trees is that respondents need to navigate through generic response categories to find their preferred response option. Under the assumption that respondents understand the underlying concepts of the generic response categories, respondents have to read fewer response options in the search tree compare to the long list. Thus, I expect that search trees facilitate respondents in giving answers, but not as much as combo boxes do, as respondents have to navigate through generic response categories.

It is important for survey practitioners to identify measurement differences between

tional qualifications, out-dated educational qualifications, and synonyms that are not covered by the other interface designs.

[4]Yet, there is no fuzzy search included in the search algorithm of the combo box. Thus, the search algorithm does not account for spelling mistakes. However, the search algorithm ignored special characters, the number of space characters as well as upper and lower cases. The suggestions were presented in alphabetical and hierarchical order.

interface designs, as measurement differences might introduce measurement error (Groves et al., 2011, p. 281). The combo box offers more flexibility in answering the education question compared to the search tree or the long list. Hence, it is more likely that respondents choose more specific educational qualifications instead of the generic terms offered by the established interface designs. Thus, I expect the combo box to produce the smallest number of consistent answers when it is compared with an established interface design of previous waves.

Another important aspect for survey organizations is whether responses can be automatically coded, as manual response coding is prone to errors and time consuming (Conrad et al., 2016). Most of the responses of the combo box should be automatically codeable. However, if answers do not match to the entries of the underlying database, then the answers need to be post-coded. Long lists and search trees offer a closed-ended response format. Thus, survey organizations get the answers in the desired response format and hence, post-coding is avoided when search trees and long lists are implemented. Due to choosing the "other" category, it is possible that responses needed to be post-coded in case of long lists and search trees. However, when respondents see the list of answer options they normally choose one of the response options, as they are less likely to optimize their response when they have to take more effort to give their exact response (e.g., answering in a follow-up question after choosing the "other" category). In general, I expect long lists and search trees to produce fewer responses for post-coding than combo boxes.

In addition to respondents' highest educational qualification, respondents were asked to report any further educational qualifications[5]. The burden to respond to a question is influenced by the interface design (Schwarz, 1999). Furthermore, response difficulty is a function of the question-specific attributes (e.g., the difficulty to investigate the list of response options, see Krosnick and Presser, 2009, p. 274). Moreover, respondent's motivation to optimize is likely to increase with response difficulty (see Krosnick and Presser, 2009). And finally, Eckman et al. (2014) showed that when respondents identify filter questions, they often misreport on these questions to avoid follow-up questions (in this case more questions on respondents' educational qualifications). I assume that respondents are more likely to misreport to filter questions when they were challenged by

[5]The question for more educational qualifications was looped after each education question. Respondents could receive the education question up to six times.

the interface design. Because I presume a lower response burden for combo boxes, I expect that respondents in the combo box condition mention more educational qualifications than in the search tree or the long list.

4.4 Study design

The split-ballot experiment was conducted in the German Socio-Economic Panel Innovative Sample (SOEP-IS), which started in 2011 (referred to as wave 1) and interviews respondents on a yearly basis (Richter and Schupp, 2015, 2012). The SOEP-IS is based on a representative sample of private households in Germany. The survey covers methodological and thematic research question, which are designed by researchers (user-designed) and implemented in the panel survey after the research proposal has been approved by the SOEP research committee (Richter and Schupp, 2015).

In 2014 (referred to as wave 2) the combo box, the search tree, and the long list were implemented in a split-ballot experiment in the SOEP-IS (three splits, for an illustration see figure 4.4). Furthermore, considered the response to the established interface design asked in 2011 (for an illustration of the study design see 4.4). In 2014[6] 5,141 respondents participated in the SOEP-IS (Richter and Schupp, 2012). However, only respondents who participated in the panel since 2011[7] (Sample I, potentially 1,278 respondents) were eligible for my experiment on interface designs, as these respondents were asked the established SOEP educational question in wave 1 (closed-ended question with 10 response options, for an illustration see figure A4.1). Furthermore, only respondents who obtained their highest qualification in Germany and who did not change their educational level in wave 2 were included in the sample (these respondents were filtered out in the survey). Moreover, I excluded respondents who had relatively fast and relatively slow response times. Due to these sample restrictions I ended up with 1,039 respondents, 349 respondents got the long list, 339 respondents got the combo box, and 351 respondents got the search tree.

Normally, respondents of the SOEP-IS are interviewed in a face-to-face interview (CAPI); however, for the purpose of this study interviewers were asked to turn the computer around so that respondents could answer the question with the interface designs

[6] doi: 10.5684/soep.is.2014
[7] doi: 10.5684/soep.is.2011

Wave 1	Wave 2
• long list with 10 response options	• long list with 28 response options
	• search tree with 38 response options
	• combo box with 400 response options

Figure 4.4: Study design.

on their own (CASI).

The question wording for all three experimental conditions was: "What is your highest educational qualification?". Respondents assigned to the long list were asked to choose their highest educational qualification from the list. Those respondents assigned to the combo box or the search tree were instructed to select the best matching result for their highest educational qualification attained.

4.5 Operationalization

This paper investigates the response burden and the data quality associated with alternative interface designs for questions with long lists of answer options. In particular, I examine two indicators of response burden - response times and response editing - and three indicators of data quality - consistency of responses given, codeability of responses, and the number of educational qualifications mentioned.

4.5.1 Response burden

Response times In this paper I use response times as a proxy for the complexity of the interface design and hence, as a proxy for response burden (see Olson and Parkhurst, 2013, p. 45; Malhotra, 2008; Turner et al., 2014; Yan and Tourangeau, 2008). Response times were captured one-side for the three experimental conditions in wave 2, but not for the established interface design asked in wave 1. Due to the skewed distribution of the response times (skewness = 14.5, kurtosis = 224.5) I trimmed the outliers by excluding the fastest and slowest 1% (21 observations, faster than 3.7 seconds and slower than 378 seconds) based on a graphical display and a sensitivity analysis (see Ratcliff,

1993). Furthermore, I took the natural logarithm to normalize the values (for methods for dealing with outliers see Ratcliff, 1993).

Response editing I examine whether response editing occurs before respondents give their final answer. Respondents change their answers because they either misread or misunderstood the question or because they accidentally clicked on the wrong response option. Hence, edited responses are another indicator for the complexity of the interface designs (see Healey, 2007; Heerwegh and Loosveldt, 2002). I captured whether respondents selected more than one response before they left the screen for the first time. However, it was not possible to capture entry changes for the experimental condition of the long list of radio buttons. Therefore, this analysis focuses on the combo box and the search tree.

4.5.2 Data quality

Consistency I operationalize data quality as consistency between the answer given in the established interface design used in wave 1[8] and the answers given in the three interface designs used in wave 2. Consistency between answers was measured by transforming the responses from the experiments into the variant of the International Standard Classification of Education (ISCED, for details see Schneider, 2013) which was used by the SOEP-IS in wave 1 (SOEP, 2014, p. 54). Thus, I transferred the very differentiated measurements of the alternative interface designs into the generic measurements (six-point scale) of the established interface design.

Codeability The paper investigates the codeability of answer options. Codeability is operationalized as whether the responses of the combo box and the search tree were automatically coded or whether a human coder was used. When a response was not automatically coded we differentiated whether the human coder was able to code the response or not. The answers were considered as not automatically codeable if "other degree" was chosen in the search tree or the long list.

[8] I harmonized respondents' educational level with their answers in the panel waves 2012 (doi: 10.5684/soep.is.2012) and 2013 (doi: 10.5684/soep.is.2013) to avoid mismatches between the interface designs due to changes in respondents' personal education history.

Number of educational qualifications The number of educational qualifications mentioned (the question was looped) were compared to see whether the interface design resulted in different numbers of reported educational qualifications.

Finally, I examine whether logarithmic response times (later referred to as logarithmic response times), editing responses and consistency between waves differ by respondent characteristics. I included mainly socio-demographic characteristics in the multivariate models, as Yan and Tourangeau (2008) showed that these variables are associated with cognitively demanding questions. For example, response times are not only affected by question characteristics, but also by respondent characteristics such as age and education (Yan and Tourangeau, 2008). Furthermore, I examine the interactions of age and educational level with the alternative interface designs to see whether response burden and data quality vary across interface designs by respondent characteristics.

4.6 Results

4.6.1 Response burden

Response times First, I examine response times across the three experimental conditions. Table 4.1 shows descriptive statistics of the response times in seconds. The combo box has the lowest response times with an average response time of 41 seconds and median response time of 32 seconds. The long list takes the longest to answer with an average response time of 50 seconds and a median response time of 43 seconds. The response times of the search tree lies in between the response times of the combo box and the research tree. The difference between the minimum and the maximum response times indicates that the amount of variation for the combo box (difference = 373) is larger than for the search tree (difference = 315) or the long list (difference = 240). These differences indicate that the combo box produces a larger variety of response times compared to the search tree and the long list. However, 75 percent of the respondents have a response time below 52 seconds which is relatively low compared to the long list were 75 percent of the respondents have a response time below 75 seconds.

Table 4.2 presents the results of the tests of difference between the interface designs for the mean and the median response times. I find a significant difference in response

Table 4.1: Response times in seconds by interface designs.

	Mean	SD	Median	Min.	Max.	25^{th} percentile	75^{th} percentile	n
Long list	50	38	43	4	244	17	75	349
Search tree	45	41	35	4	319	22	54	351
Combo box	41	38	32	4	377	20	52	339
Overall	45	39	34	4	377	20	61	1,039

NOTE. – SD = standard deviation, Min.= Minimum, Max.= Maximum, n = number of observations.
21 observations were excluded because they were in the 1st and the 99th percentile.

times between the three alternative interface designs ($F_{2;1,039} = 4.6$, $p = 0.01$). The difference in mean response times is 5 seconds between the long list and the search tree, the difference is not significant. While the difference of 9 seconds in the mean response time between the long list and the combo box is significant (t = -3.03, $p = 0.01$), the difference of 4 seconds in the mean response time between the combo box and the search tree is not significant.

When comparing the median response times between interface designs, I find no significant difference between the long list and the search tree. However, I detect a significant difference in median response time between the long list and the combo box ($\chi^2 = 4.89$, $d.f. = 1$, $p = 0.03$). By the same token, I observe a significant difference in median response times between the search tree and the combo box ($\chi^2 = 3.07$, $d.f. = 1$, $p = 0.08$).

Table 4.2: Tests of difference between interface designs for response times.

	Tests of difference for means[1]	Tests of difference for medians[2]
Long list vs. search tree	$t = -1.56, p = 0.36$	$\chi^2 = 1.65, d.f. = 1, p = 0.12$
Long list vs. combo box	$t = -3.03, p = 0.01$	$\chi^2 = 4.89, d.f. = 1, p = 0.03$
Search tree vs. combo box	$t = 1.49, p = 0.41$	$\chi^2 = 3.07, d.f. = 1, p = 0.08$

NOTE. – 21 observations were excluded because they were in the 1st or the 99th percentile.
[1] Tests of difference are based on a pairwise comparison of means with Bonferroni's correction.
[2] Tests of difference are based on a median χ^2 test.

The linear model in table 4.3 reports the impact of respondent characteristics on logarithmic response times in seconds[9]. I find answer consistency between panel waves

[9]The estimates are also robust when I estimated fixed-effects models to account for respondents being

1 and 2 to be significantly negatively associated with logarithmic response times. Conversely, medium vocational qualification and higher vocational qualification are positively associated with logarithmic response times. I observe no significant association of response times with age, age squared, and the usage of the combo box or the search tree.

Table 4.3: Linear regression of respondents' characteristics and interface designs on the natural logarithm of response times in seconds.

	$\hat{\beta}$	Std. err.
Age	0.01	0.02
Age^2	−0.00	0.00
Being female	−0.06	0.05
Ref. General elementary qualification		
Medium vocational qualification	0.39*	0.17
Vocational qualification and Abitur	0.12	0.34
Higher vocational qualification	0.45^{+}	0.25
Higher educational qualification	0.29	0.19
Answer consistency	−0.23***	0.07
Ref. Long list		
Combo box	−0.01	0.45
Search tree	−0.54	0.45
*Ref. Educational qualification*long list*		
Medium vocational*combo box	−0.43*	0.20
Medium vocational*search tree	0.36	0.22
Vocational and Abitur*combo box	−0.26	0.40
Vocational and Abitur*search tree	0.38	0.39
Higher vocational*combo box	−0.78**	0.29
Higher vocational*search tree	0.34	0.34
Higher educational*combo box	−0.54**	0.21
Higher educational*search tree	0.17	0.25
Number of respondents	1,039	

NOTE. – $\hat{\beta}$ = coefficients, Std. err. = standard errors, ref. = Reference category
Std. err. adjusted for 136 interviewer clusters.
$^{+}$ p < 0.10, * p < 0.05, ** p < 0.01, *** p < 0.001

Table 4.3 indicates that the interaction effects of educational qualifications and in-

clustered in interviewers.

terface designs are significant in three cases. For example, respondents with medium vocational qualification, higher vocational qualification, and higher educational qualification who used the combo box have a significantly negative association with logarithmic response times compared to respondents with the same educational qualifications who used a long list. Furthermore, tests of difference show that respondents with medium vocational qualification, higher vocational qualification, and higher educational qualifications using a combo box respond significantly faster compared to the respondents with the same educational levels using a search tree ($F_{2;134} = 7.78$, $p = 0.00$; $F_{2;134} = 10.49$, $p = 0.00$; $F_{2;134} = 2.29$, $p = 0.06$). These findings indicate that respondents with vocational qualifications need less time to respond using a combo box than using a long list or a search tree.

Response editing Second, I investigate whether there are differences in editing responses between the search tree and the combo box. Figure 4.5 shows the percentage of edited responses for the two experimental conditions. The results demonstrate that 41 percent of the respondents edited their response when using the combo box and 51 percent of the respondents edited their response when using the search tree. The difference is significant between the two experimental conditions ($t = 2.72$, $p = 0.01$).

Table 4.4 gives the results of a logistic regression on editing responses under the control of respondent characteristics. I find none of the main effects of the sociodemographic variables to be significant, besides medium vocational and higher educational qualification. Respondents with medium vocational qualification and higher educational qualification edited their responses significantly less than respondents with other educational qualifications. When answers between wave 1 and wave 2 were consistent respondents are significantly more likely to change their answer than respondents who gave inconsistent answers between wave 1 and wave 2. Furthermore, I detect a significant increase of the logarithmic response times when responses were edited.

The results reveal that respondents edited their responses significantly more when using the search tree compared to the combo box; however, older respondents seem to edit their responses less often when using a search tree than using the combo box. The interaction of the search tree condition and medium vocational qualification is significantly negative indicating that respondents with a medium vocational qualification edited their responses less using a search tree than respondents with the same educational

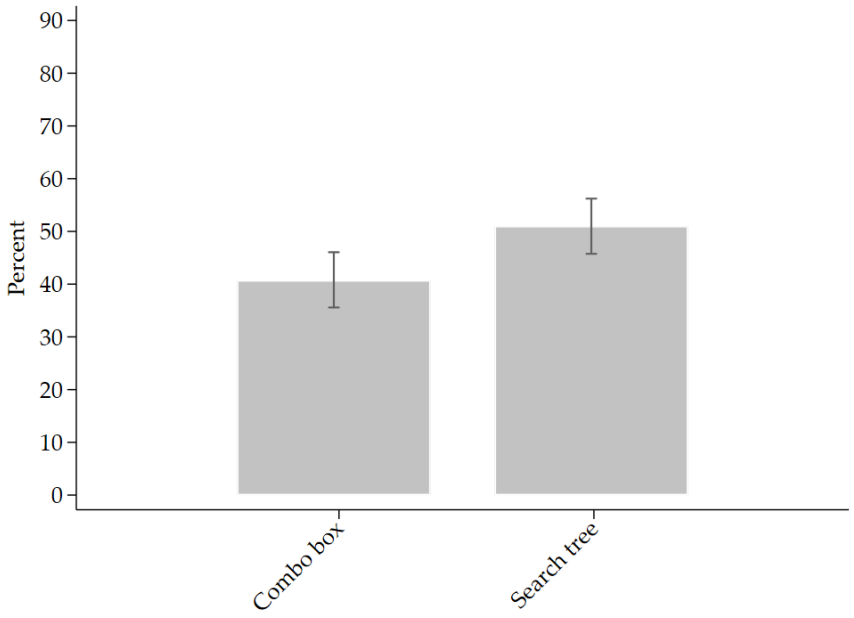


Figure 4.5: Percentage of edited responses between interface designs of wave 2 with 95% confidence intervals.

qualification who used the combo box. However, respondents with higher educational qualifications edited responses significantly more in the search tree condition than in the combo box.

4.6.2 Data quality

Consistency To investigate data quality across interface designs I compare the consistent answers of the responses given in wave 1 with the responses given in wave 2. Figure 4.6 presents the percentage of answer consistency between the long list used in wave 1 and the three interface designs used in wave 2. All three interface designs achieve an answer consistency over 80 percent. As indicated in figure 4.6 there is no significant difference in answer consistency between the three experimental conditions (pairwise comparisons of means with Bonferroni's correction: combo box vs. long list t = -1.58, p = 0.35, search tree vs. long list t = -0.96, p = 1.00, search tree vs. combo box t = 0.63, p = 1.00).

To investigate whether the answer consistency differs by respondent characteristics I

Table 4.4: Logistic regression of respondents' characteristics and interface designs on edited responses.

	$\hat{\beta}$	Std. err.
Age	0.02	0.04
Age^2	−0.00	0.00
Being female	0.17	0.16
Ref. General elementary qualification		
Medium vocational qualification	−0.81**	0.30
Vocational qualification and Abitur	−0.41	0.52
Higher vocational qualification	−0.21	0.57
Higher educational qualification	−1.00*	0.42
Answer consistency	0.63*	0.27
Log. response times	0.43***	0.13
Ref. Combo box		
Search tree	0.49***	0.12
*Ref. Combo box*age*		
Search tree*age	−0.16**	0.05
Ref. Combo boxage^2*		
Search tree*age^2	0.00**	0.00
*Ref. Educational qualification*combo box*		
Medium vocational*search tree	−1.21**	0.38
Vocational and Abitur*search tree	−0.87	0.81
Higher vocational*search tree	−0.93	0.76
Higher educational*search tree	1.01^{+}	0.60
Number of respondents	689	

NOTE. – $\hat{\beta}$ = coefficients, Std. err. = standard errors
Std. err. adjusted for 125 interviewer clusters.
$^{+}$ p < 0.10, * p < 0.05, ** p < 0.01, *** p < 0.001

estimated a logistic regression model with socio-demographic covariates (see table 4.5). Age, age squared, medium vocational qualifications, higher educational qualifications, and logarithmized response times affect the answer consistency significantly. With increasing age, the answer consistency initially increases, but in higher ages, it decreases again. Respondents with medium vocational qualifications and higher educational qualifications have significantly more consistent answers than respondents with general elementary qualifications. Faster respondents produce less consistent answers. However,

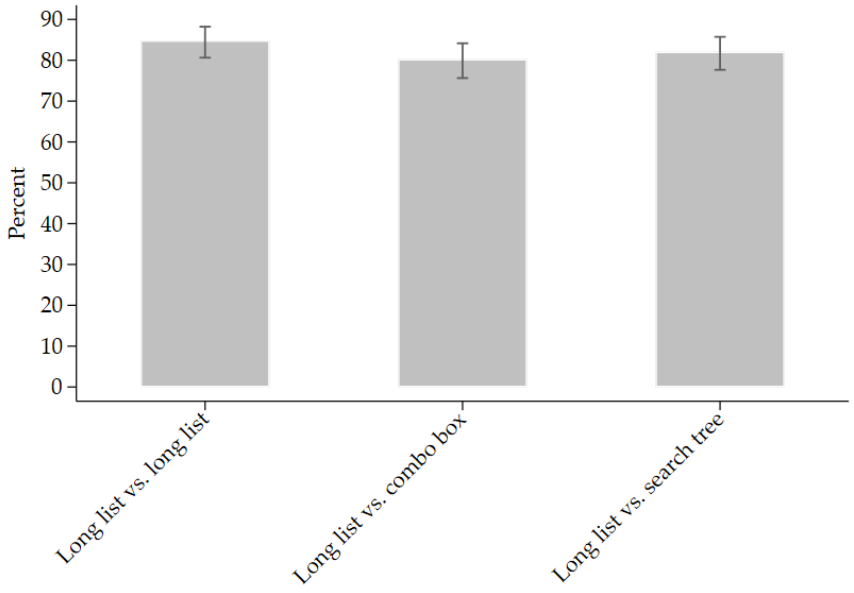


Figure 4.6: Percentage of answer consistency between the established interface design used in wave 1 and the alternative interface designs used in wave 2 with 95% confidence intervals.

the interface designs are not significantly associated with answer consistency. In addition, I find no interaction effects between age, age squared, or educational qualifications with the interface designs to be significant (not reported). This suggests that the differences in answer consistencies across experimental conditions do not vary by respondent characteristics.

Codeability Table 4.6 illustrates the codeability of the highest educational qualification mentioned in percentage. When using the long list, 100 percent of the responses were codeable, as all respondents found their highest educational qualification in the list of answer options. In the case of the combo box 52 percent of the responses were automatically coded, meaning that the entered response was found in the underlying database. The database had no suitable match in 38 percent of the responses entered in the combo box, but human coders identified a code[10]. In 10 percent of the cases no coding of the entered response was possible as the response format was unintended, for example, the text field was empty (1 case), or the respondents did not give their highest

[10]These human coded responses were included in the database for later use.

Table 4.5: Logistic regression of respondents' characteristics and interface design on answer consistency between wave 1 and wave 2.

	$\hat{\beta}$	Std. err.
Age	0.08*	0.04
Age^2	-0.00^{+}	0.00
Being female	0.21	0.17
Ref. General elementary qualification		
Medium vocational qualification	1.70***	0.47
Vocational qualification and Abitur	−0.33	0.67
Higher vocational qualification	0.28	0.70
Higher education qualification	1.73**	0.65
Log response times	−0.43***	0.13
Ref. Long list		
Combo box	0.04	1.53
Search tree	1.34	1.28
Number of respondents	1,037	

NOTE. – $\hat{\beta}$ = coefficients, Std. err. = standard errors, ref. = Reference category
Std. err. adjusted for 136 interviewer clusters.
2 cases were omitted because there were not enough cases on the value "inadequate" on the educational level.
Controlled for interactions of the experimental conditions with age, age squared, educational level.
$^{+}$ p < 0.10, * p < 0.05, ** p < 0.01, *** p < 0.001

educational qualification (e.g., a lower education). In case of the search tree 95 percent of the responses were coded automatically. However, 5 percent of the respondents could not identify their educational qualification in the search tree and answered a follow-up question with a standard text field. Out of these responses, 2 percent were codeable by human coders, whereas 3 percent of the respondents entered a text which was in an unintended response format (e.g., occupations). In total, I find significant variation between the three interface designs in terms of response codeability ($\chi^2 = 741.83$, $d.f. = 4$, $p = 0.00$).

Number of educational qualifications Table 4.6 indicates that 88 percent of the respondents in the long list condition mentioned one educational qualification. This finding is in contrast to the combo box and the search tree condition where 81 percent and 80 percent respectively, indicated to have one educational qualification. In the experi-

Table 4.6: Codeability and number of educational qualifications mentioned by interface designs in percentage.

	Long list	Combo box	Search tree
Codeability			
Automatically coded	0	52	95
Codeable	100	38	2
Not codeable[1]	0	10	3
Number of educational qualifications[2]			
1 qualification	88	81	80
2 qualifications	10	15	15
$\geq$ 3 qualifications	2	4	5
Number of respondents	349	339	351

NOTE. – [1] Not codeable is defined as an unintended format of the response.
[2] Only valid educational qualifications were considered.

mental condition of the long list, 10 percent mentioned a second educational qualification, whereas 15 percent said that they had a second educational qualification in the combo box condition and the search tree condition. Finally, 2 percent of the long list users, 4 percent of the search tree users, and 5 percent of the combo box users indicated to have three or more educational qualifications. The difference across interface designs is significant ($\chi^2 = 9.25$, $d.f. = 4$, $p = 0.06$). Consequently, respondents mention a similar number of educational qualifications when using the combo box or the search tree and mention less educational qualifications when using the long list.

4.7 Discussion

It is unfortunate that we know so little about the optimal design of long-list questions in computer-assisted surveys. This article contributes to this research gap, by evaluating the differences in response burden and data quality between two alternative interface designs for long-list questions. For this purpose, I conducted a split-ballot experiment in a computer-assisted panel study in Germany on a key socio-economic background question: the highest educational qualification obtained.

To examine whether a long list of radio buttons, a combo box or a search tree best

facilitates the answering of long-list questions, I compared response times, response editing, answer consistency, codeability of responses, and the number of educational qualifications mentioned. I observed significantly lower response times when using the combo box compared to the search tree or the long list. However, when controlling for respondents' socio-demographics, the effect of the interface designs diminishes. Respondents with medium vocational, higher vocational and higher educational qualifications are faster when using the combo box than respondents with the same educational qualifications using the long list. Thus, the interface design effect is moderated by the respondents' educational level. Furthermore, answer consistency is associated with lower response times, i.e., respondents that are very sure of which answer they want to choose to respond faster independently of the interface design. These findings indicate that the combo box reduces the response burden for some groups of respondents.

The split-ballot experiment revealed that there is a significant difference in response editing between the search tree and the combo box. Controlling for respondent characteristics, I find that the search tree results in more response editing than the combo box. The exceptions are respondents with medium vocational qualifications and higher vocational qualifications, who edited their responses less in the search tree condition compared to the combo box condition. The results indicate that with increasing age the search tree is associated with less response editing compared to the combo box. Overall, it seems that the combo box facilitates most respondent groups better in answering long-list questions than the search tree.

I compared the answer consistency between the established long list used in wave 1 and the three interface designs used in wave 2. I find no significant difference between the combo box, the search tree, and the long list. Although respondents with medium educational and higher educational qualifications have a higher answer consistency than respondents with general educational qualifications, I find no evidence that consistency varies by educational qualification across the three experimental conditions. There is no evidence that long lists, combo boxes, or search trees generate different levels of data quality.

Researchers aim to avoid unintended answers of respondents, as these answers are mostly resulting in post-coding or item nonresponse. The results reveal that 100 percent of the responses of the long list, 90 percent of the responses of the combo box, and 97 percent of the responses of the search tree produced codeable values. However, only

52 percent of the responses in the combo box were automatically coded. The other 38 percent of the answers needed a post-survey coding, mainly because the entry did not exist in the underlying database or because of typos (Schneider et al., 2016a). These findings suggest that researchers have to put more effort in post-survey coding when using the combo box compared to the long list or the search tree.

Finally, I examined the number of educational qualifications mentioned by the respondents. I find that respondents in the combo box and the search tree condition mention more educational qualifications than respondents in the experimental condition of the long list. This finding is important for researchers who are interested in measuring not only the highest educational qualification, as respondents seem to mention more than one educational qualification in the combo box and the search tree condition.

4.8 Conclusions

Survey research aims to facilitate respondents in finding the optimal answer and in enhancing the quality of measurements. Alternative interface designs for long-list questions are already implemented in surveys, yet the impact of these interface designs on response burden and measurement quality is unknown. This study represents one step in improving interface designs of long-list questions, as I compared two alternative interface designs - combo box and search tree - with the established way of measuring questions with a large number of response options - a long list of radio buttons.

This study suggests that combo boxes are an alternative to the established interface design of long-list questions, as combo boxes have a lower response burden without having a lower answer consistency than search trees or long lists. Search trees have a higher response burden than combo boxes; however, respondents are faster in answering search trees than long lists. In addition, the combo box and the search tree offer more granularity for the example of educational qualifications (Schneider et al., 2016a), as more response options can be shown than in the established long lists. Nevertheless, one cannot ignore the effort researchers have to put in post-survey coding when using alternative interface designs. Yet, neither the search tree nor the combo box is the silver bullet for the design of long-list questions and researchers have to decide for one or the other interface design based on their research intentions. For example, I would

recommend using a combo box when high measurement granularity is needed. But if the answer needs to be coded on the spot, a search tree works best. Currently, there still is a trade-off between a decrease in the response burden and an increase of post-survey coding.

For the analysis presented in this paper, answer consistency between panel waves was used as a proxy indicator of data quality. Future research would benefit from a comparison of the measurements of combo boxes and search trees with true values. Furthermore, no information on whether interviewers intervened or helped respondents in filling out the questions in the CASI conditions was available. Therefore, I used clustered standard errors to considered interviewers as a source of variation. A next step would be to include interviewer debriefing questions to get more insights into the interview situation. The same is true for investigating whether respondents' familiarity with technology reduces response times. Furthermore, research on combo boxes could improve the search algorithm for the database request (e.g., fuzzy matching), which might help respondents provide more codeable answers.

There are new machine learning algorithms which could be implemented in the survey procedure, which might increase the amount of automatically coded responses (for example Schierholz et al., 2017). Yet, we do not know if this approach is more promising than database lookups with regard to data quality and usability. Thus, a comparison of interface designs with different back-ends - database lookups versus machine learning suggestions - would be an interesting extension of the existing research.

Despite these limitations, the results are encouraging for those who want to implement combo boxes or search trees in their survey procedure. Likewise, if these interface designs worked with the example of education, they should be suitable for different kinds of questions with many answer options, such as prescription drugs, occupations, company names or brands. My evaluation of two alternative interface designs contributes to research on the optimal interface design for long-list questions and shows how survey practitioners can improve the measurement of a key socio-demographic variable, namely educational qualifications.

Acknowledgments

This work was supported by the grant SAW-2013-GESIS-5 249 provided by the Leibniz-Association through the Leibniz competition. I gratefully acknowledge support from the SOEP Innovation Sample and Collaborative Research Center (SFB) 884 "Political Economy of Reforms" at the University of Mannheim. I especially thank Silke Schneider, Verena Ortmanns, Beatrice Rammstedt and David Richter for their support in making this experiment possible.

References

Christian, L. M. and Dillman, D. A. (2004). The Influence of Graphical and Symbolic Language Manipulations on Responses to Self-Administered Questions. *Public Opinion Quarterly*, 68(1):57–80.

Conrad, F. G., Couper, M. P., and Sakshaug, J. W. (2016). Classifying Open-Ended Reports: Factors Affecting the Reliability of Occupation Codes. *Journal of Official Statistics*, 32(1):75–92.

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

Couper, M. P., Kennedy, C., Conrad, F. G., and Tourangeau, R. (2011). Designing Input Fields for Non-Narrative Open-Ended Responses in Web Surveys. *Journal of Official Statistics*, 27(1):65–85.

Couper, M. P., Tourangeau, R., Conrad, F. G., and Crawford, S. D. (2004). What They See Is What We Get: Response Options for Web Surveys. *Social Science Computer Review*, 22(1):111–127.

Couper, M. P. and Zhang, C. (2016). Helping Respondents Provide Good Answers in Web Surveys. *Survey Research Methods*, 10(1):49–64.

de Leeuw, E. D. (2005). To Mix or Not to Mix Data Collection Modes in Surveys. *Journal of Official Statistics*, 21(2):233–255.

Dillman, D. A., Phelps, G., Tortora, R., Swift, K., Kohrell, J., Berck, J., and Messer, B. L. (2009). Response Rate and Measurement Differences in Mixed-Mode Surveys using Mail, Telephone, Interactive Voice Response (IVR) and the Internet. *Social Science Research*, 38(1):1–18.

Eckman, S., Kreuter, F., Kirchner, A., Jäckle, A., Tourangeau, R., and Presser, S. (2014). Assessing the Mechanisms of Misreporting to Filter Questions in Surveys. *Public Opinion Quarterly*, 78(3):721–733.

Fowler, F. J. (1995). *Improving Survey Questions: Design and Evaluation*, volume 38 of *Applied Social Research Methods Series*. Sage Publications, Thousand Oaks, London, New Dehli.

Funke, F. and Reips, U.-D. (2007). *Dynamic Form: Online Surveys 2.0*. Paper presented at the General Online Research Conference (GOR 2007), Leipzig, Germany.

Groves, R. M., Fowler, F. J., Couper, M. P., Lepkowski, J. M., Singer, E., and Tourangeau, R. (2011). *Survey Methodology*, volume 561 of *Wiley Series in Survey Methodology*. John Wiley & Sons, Hoboken, NJ, 2nd edition.

Healey, B. (2007). Drop Downs and Scroll Mice: The Effect of Response Option Format and Input Mechanism Employed on Data Quality in Web Surveys. *Social Science Computer Review*, 25(1):111–128.

Heerwegh, D. and Loosveldt, G. (2002). An Evaluation of the Effect of Response Formats on Data Quality in Web Surveys. *Social Science Computer Review*, 20(4):471–484.

Keusch, F. (2014). The Influence of Answer Box Format on Response Behavior on List-Style Open-Ended Questions. *Journal of Survey Statistics and Methodology*, 2(3):305–322.

Krosnick, J. A. and Alwin, D. F. (1987). An Evaluation of a Cognitive Theory of Response-Order Effects in Survey Measurement. *Public Opinion Quarterly*, 51(2):201–219.

Krosnick, J. A. and Presser, S. (2009). Question and Questionnaire Design. In Wright, J. D. and Marsden, P. V., editors, *Handbook of Survey Research*, chapter 9, pages 263–315. Elsevier, San Diego, CA.

Lenzner, T., Kaczmirek, L., and Lenzner, A. (2010). Cognitive Burden of Survey Questions and Response Times: A Psycholinguistic Experiment. *Applied Cognitive Psychology*, 24(7):1003–1020.

Malhotra, N. (2008). Completion Time and Response Order Effects in Web Surveys. *Public Opinion Quarterly*, 72(5):914–934.

Olson, K. and Parkhurst, B. (2013). Collecting Paradata for Measurement Error Evaluations. In Kreuter, F., editor, *Improving Surveys with Paradata: Analytic Uses of Process Information*, volume 581 of *Wiley Series in Survey Methodology*, chapter 3, pages 43–72. John Wiley & Sons, Hoboken, NJ.

Ratcliff, R. (1993). Methods for Dealing With Reaction Time Outliers. *Psychological Bulletin*, 114(3):510–532.

Redline, C. D. and Dillman, D. A. (2002). The Influence of Alternative Visual Designs on Respondents' Performance With Branching Instructions in Self-Administered Questionnaires. In Groves, R. M., Dillman, D. A., Eltinge, J. L., and Little, R. J. A., editors, *Survey Nonresponse*, chapter 12, pages 179–196. Wiley & Sons, New York, NY.

Redline, C. D., Tourangeau, R., Couper, M. P., Conrad, F. G., and Ye, C. (2009). The Effects of Grouping Response Options in Factual Questions with Many Options. In *JPSM Research Paper*, Washington, DC. Annual Conference of the Federal Committee on Statistical Methodology.

Richter, D. and Schupp, J. (2012). *SOEP Innovation Sample (SOEP-IS) – Description, Structure and Documentation*, volume 463 of *SOEP Papers on Multidisciplinary Panel Data Research*. Deutsches Institut für Wirtschaftsforschung, DIW, Berlin, DE.

Richter, D. and Schupp, J. (2015). The SOEP Innovation Sample (SOEP IS). *Schmollers Jahrbuch*, 135(3):389–399.

Schierholz, M., Gensicke, M., Tschersich, N., and Kreuter, F. (2017). Occupation Coding During the Interview. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 181(2):379–407.

Schneider, S. L. (2008). Suggestions for the Cross-National Measurement of Educational Attainment: Refining the ISCED-97 and Improving Data Collection and Coding Procedures. In Schneider, S. L., editor, *The International Standard Classification of Education (ISCED-97). An Evaluation of Content and Criterion Validity for 15 European Countries*, chapter 17, pages 311–330. MZES, Mannheim, DE.

Schneider, S. L. (2013). The International Standard Classification of Education 2011. In Birkelund, G. E., editor, *Class and Stratification Analysis*, volume 30 of *Comparative Social Research*, pages 365–379. Emerald Group Publishing Limited, Bradford, UK.

Schneider, S. L., Briceno-Rosas, R., Herzing, J. M. E., and Ortmanns, V. (2016a). Overcoming the Shortcomings of Long List Showcards: Measuring Education With an Adaptive Database Lookup. In *9th International Conference on Social Science Methodology*, RC33 Conference, Leicester, UK.

Schneider, S. L., Joye, D., and Wolf, C. (2016b). When Translation is not Enough: Background Variables in Comparative Surveys. In Wolf, C., Joye, D., Smith, T. W., and Fu, Y.-C., editors, *The SAGE Handbook of Survey Methodology*, chapter 20, pages 288–307. Sage, London, UK.

Schwarz, N. (1999). Self-Reports: How the Questions Shape the Answers. *American Psychologist*, 54(2):93–105.

Smith, T. W. (1995). Some Aspects of Measuring Education. *Social Science Research*, 24(3):215–242.

Smyth, J. D., Dillman, D. A., Christian, L. M., and Stern, M. J. (2006). Effects of Using Visual Design Principles to Group Response Options in Web Surveys. *International Journal of Internet Science*, 1(1):6–16.

SOEP, G. (2014). *SOEP 2013 – Documentation of Person-Related Status and Generated Variables in PGEN for SOEP v30*. SOEP Survey Paper 250 (Series D). DIW/SOEP, Berlin, DE.

Stern, M. J. (2008). The Use of Client-Side Paradata in Analyzing the Effects of Visual Layout on Changing Responses in Web Surveys. *Field Methods*, 20(4):377–398.

Tijdens, K. (2015). Self-Identification of Occupation in Web Surveys: Requirements for Search Trees and Look-up Tables. *Survey Methods: Insights from the Field*. https://doi.org/10.13094/SMIF-2015-00008.

Tourangeau, R., Groves, R. M., and Redline, C. D. (2010). Sensitive Topics and Reluctant Respondents: Demonstrating a Link between Nonresponse Bias and Measurement Error. *Public Opinion Quarterly*, 74(3):413–432.

Turner, G., Sturgis, P., and Martin, D. (2014). Can Response Latencies Be Used to Detect Survey Satisficing on Cognitively Demanding Questions? *Journal of Survey Statistics and Methodology*, 3(1):89–108.

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

Appendix

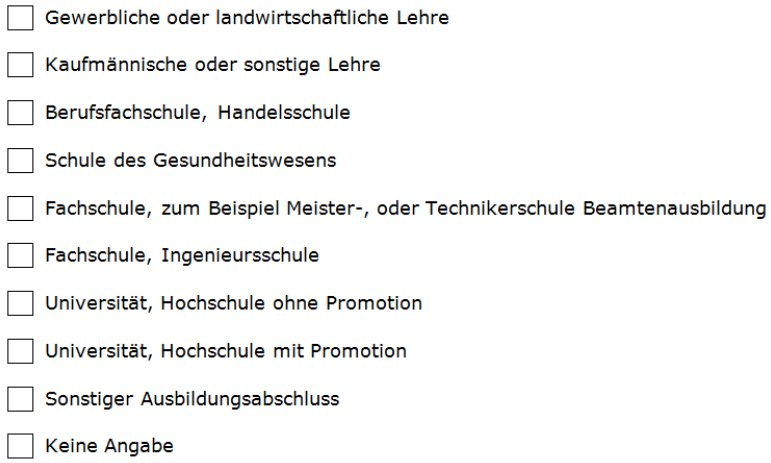


Figure A4.1: Established interface design - Long list with 10 items.

Chapter 5

Paper IV

How do respondents use combo boxes? An evaluation

Abstract

Although survey practitioners implement combo boxes already, it is yet unknown how respondents use combo boxes and which response difficulties emerge when respondents use combo boxes. For this purpose, we conducted an eye-tracking study in combination with cognitive interviews to test the usability of combo boxes when asking the question on the highest educational qualification. Our results indicate that respondents may not know that they can type a response rather than select a response option and vice versa. Hence, respondents either use combo boxes as an open-ended or as a closed-ended question. We find some response difficulties when using a combo box which are normally associated with the use of open-ended or closed-ended questions. We detect that the search algorithm of the combo box needs to be optimized with regard to the order of displayed response options. Based on cognitive interviews we derive which visual and verbal cues are missing in the interface designs to increase the respondents' understanding of how to use combo boxes.

How do respondents use combo boxes? An evaluation[1]

5.1 Introduction

The design of questions with a large number of response options has challenged survey practitioners of computer-assisted surveys, as survey practitioners have to decide whether they design long-list questions as an open-ended (respondents answer in their own words in standard text fields) or as a closed-ended (respondents select an answer from a choice set) question. To decide for one or the other interface design, survey practitioners have to consider advantages and disadvantages of both the open-ended and the closed-ended question format.

One advantage of open-ended question formats is that these formats do not limit respondents' answers (Foddy, 1993, p. 127). However, open-ended questions suffer from more item nonresponse than closed-ended questions (Reja et al., 2003). Furthermore, open-ended question formats require extensive coding (Fowler, 1995, p. 59) which requires the development of a coding scheme, its application by more than one coder, and a high level of agreement between human coders (Krosnick and Presser, 2009, p. 269). The high costs of these coding procedures coupled with potential errors introduced by human coders and high item nonresponse rates are the main reasons why closed-ended questions are widely-used.

One advantage of closed-ended question formats is that no coding or less coding is needed. Furthermore, respondents are guided by the response options offered (see Tourangeau et al., 2000, p. 203). However, questions in closed-ended formats limit the number of response categories to a relatively small number of response options (Fowler, 1995) which is disadvantageous when there is a large number of possible answer options. Hence, in long-list questions, researchers have to limit the number of response options in a way that the response options fit on one page/screen. In addition, the number of response options presented should not exceed respondents' willingness to read the list of response options (see Lenzner et al., 2010). Furthermore, response order effects are a common problem of closed-ended questions (Krosnick and Alwin, 1987). Thus, both open-ended and closed-ended questions have advantages and disadvantages

[1] This chapter is co-authored with Cornelia E. Neuert.

which survey researchers have to consider when deciding on one or the other question format.

With technical advancements, survey organizations developed alternative interface designs for long-list questions to assist respondents in answering long-list questions and to increase measurement quality (Callegaro et al., 2015; Couper, 2000; Tourangeau et al., 2013, ch. 6). One of these alternative interface designs are combo boxes which combine open-ended and closed-ended question formats. Combo boxes combine standard text fields with drop-down boxes. Due to this combination of question formats, respondents can either type their answer directly in the text field or select a response from the drop-down box. On the one hand, combo boxes, do not restrict respondents in their answers, while on the other hand respondents are guided by the displayed response suggestions which might reduce unintended answers (from a researchers' perspective). Thus, combo boxes combine the advantages of both open-ended and closed-ended question formats.

Because of these design advantages survey practitioners already use combo boxes for questions on occupation, prescription drugs, field of study and educational qualifications (for examples see Couper and Zhang, 2016; Tijdens, 2015), albeit the evaluations of combo boxes have been ambiguous with regard to usability and data quality (see Couper and Zhang, 2016; Tijdens, 2015). In addition, Couper and Zhang (2016) indicate that respondents may not know how combo boxes function, for instance, that they can type a response rather than select a response option from the list and vice versa. This nescience of respondents can result in more uncodeable answers when respondents use combo boxes (see Couper and Zhang, 2016). Thus, knowing whether respondents use a combo box like a standard text field or like a drop-down box helps to identify how response difficulties emerge when respondents use this alternative interface design.

This article investigates how respondents use combo boxes. Furthermore, we examine whether using combo boxes as an open-ended or as a closed-ended question format comes along with the same response difficulties or with response difficulties which are normally associated with open-ended and closed-ended question formats. For this purpose, we track eye movements to examine respondents' attention while answering the question on their highest educational qualification with the help of a combo box. Combining eye-tracking with cognitive interviews allows us to probe how respondents understood the usage of the combo box and whether they had problems when answering

the question. From the cognitive interviews, we derive recommendations on how to improve the visual presentation to clarify the functionality of combo boxes. This research adds to the literature on alternative interface designs for long-list questions, as it is only reasonable to implement combo boxes instead of long lists when respondents can use this alternative interface design properly.

5.2 Background

The question-answering process proposed by Tourangeau et al. (2000) was extended by Jenkins and Dillman (1997) and includes five steps in its simplest form: (1) respondents perceive the visual elements (e.g., graphic paralanguage, numeric and symbolic cues), (2) respondents comprehend the question, (3) respondents recall information, (4) respondents make a judgment, and finally (5) respondents give a response. In each of these steps response effects can emerge, for instance, respondents might misinterpret the visual elements, misinterpret the question, forget important information, make the wrong inference based on their accessibility, or map their intended answer onto an inappropriate response option (see Tourangeau et al., 2000, p. 8). Thus, both the verbal language and the visual presentation influence respondents' question comprehension and hence, how respondents process and respond to a survey question (Redline et al., 2009; Schwarz and Sudman, 1996; Smyth et al., 2006).

From previous research we know, that how respondents perceive visual elements of a survey question influences the measurement (see Couper, 2000; Couper et al., 2004; Schwarz, 1999; Tourangeau et al., 2004). For instance, Tourangeau et al. (2004) showed that the mid-point of rating scales varies with the spacing of the response options. Furthermore, the perception of visual cues can result in measurement variation between respondents, as respondents differ in seeing visual cues and respondents differ in their expectations (or knowledge) about visual cues (Jenkins and Dillman, 1997). For example, a scrollbar paired with a drop-down box indicates that there is further information which is not yet displayed. However, not all respondents might see the scrollbar or respondents might expect an arrow instead of a scrollbar as an indicator for further information. Differences in understanding these visual cues may cause that some respondents do not investigate the full list of response options. Thus, the visual presentation can influence

the measurement and how respondents process and respond to a survey question (see Couper et al., 2011; Smyth et al., 2009).

To reduce measurement error, respondents should understand the question as the researcher intended and all respondents should understand the question in the same way (Fowler, 1995; Presser et al., 2004). The same is true for visual cues, all respondents should understand the visual cues of a question in the same way and as the researcher intended to avoid measurement error. Hence, inadequately designed question formats may increase measurement error, when respondents cannot answer in the appropriate manner (see Reips, 2000). Therefore, it is prudent to investigate whether the visual cues of a combo box support respondents' understanding of how to use this alternative interface design.

In the following, we examine how respondents perceive the visual presentation of a combo box and how respondents make their response given a combo box. For this purpose, this study investigates whether respondents perceive combo boxes differently and as a consequence, whether their understanding of how to use a combo box differs. Furthermore, we investigate whether a difference in understanding the functions of the combo box results in specific response difficulties. By understanding how respondents use combo boxes, survey practitioners gain knowledge about how to optimize the design of combo boxes in order to decrease potential response difficulties.

5.3 Use of combo boxes

Combo boxes combine drop-down boxes and standard text fields. At first glance, combo boxes look like a standard text field (for an illustration see figure 5.1). The moment the respondent starts typing, multiple suggestions[2] emerge below the initial text box, similar to a drop-down box (for an illustration see figure 5.2).

Each additional letter reduces the number of response suggestions. There is no restriction on the maximum number of presented response suggestions; however, the presentation of response suggestions is limited to the screen size (the screen size is the same for all respondents in our study design). In case the list of response suggestions

[2] In our case, the response options are in alphabetical and hierarchical ascending order. The combo box was designed in a Blaise environment with the help of JavaScript (for further information see Schneider et al., 2018, or www.surveycodings.org/education).

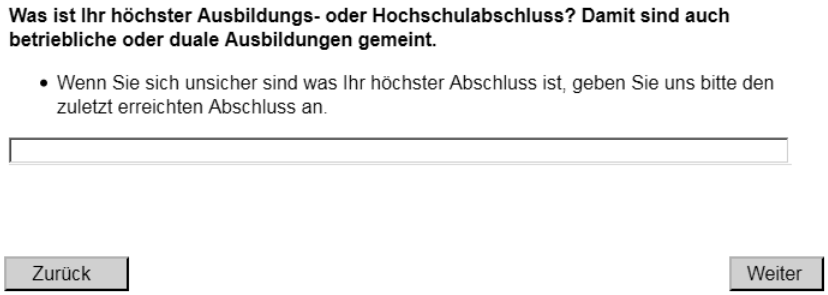


Figure 5.1: Initial interface design of the combo box.

Figure 5.2: Interface design of the combo box the moment respondents start typing.

exceeds the screen size, arrows, and a scrollbar appear on the right side of the drop-down list.

As combo boxes are a combination of open-ended and closed-ended question formats respondents can either use the combo box as a standard text field (type the response in

the text box) or as a drop-down box (start typing and choose from the list of response options). Yet, it is unknown whether respondents know that they can type a response into the text field or select one of the response suggestions and vice versa. It is also unclear whether and how respondents examine the list of response options.

There are five steps to the question-answering process all respondents go through when answering a question with a combo box. First, all respondents perceive the visual cues to identify the question itself. Next, respondents need to comprehend the questions and instructions. Then, respondents retrieve the necessary information from their memories. Subsequently, respondents estimate the relevance and completeness of their retrieved information. Afterward, respondents make their response by starting to type the response into the text field of the combo box. During the next steps of the question-answering process differences on how to use the combo boxes might occur which can be attributed to respondents' expectations about visual cues.

Use the combo box closed-ended One usage scenario is that respondents see the drop-down box which pops up below the text field when they start typing. Furthermore, they perceive that scrolling within the drop-down box is possible. These respondents examine the response suggestions and select one of the options. Hence, respondents interpret the visual cues of the combo box and come to the conclusion to use the combo box as a drop-down box. Here it is not clear whether respondents are aware of being able to type in the response.

Use the combo box open-ended Another usage scenario is that respondents enter a response, see the drop-down box, and examine the response suggestions. However, it is also possible that they do not see the drop-down box at all. In both cases, respondents type their response into the text field and press the submit button without selecting a response option from the list of response suggestions because they perceived and interpreted the visual cue to use the text field as a standard text field. In this case, it is unclear whether respondents did know that they also could have clicked on a response suggestion.

5.4 Methods

5.4.1 Measuring educational qualification with combo boxes

We illustrate the usage of combo boxes by investigating the question on the highest educational qualification[3]. Asking about the highest educational qualification with a combo box is particularly suitable because it allows respondents to report their educational qualification in their own words rather than forcing them to choose from a limited number of fixed response categories, which hardly reflect continuously changing educational systems or foreign qualifications (for more information on measuring educational qualifications see Schneider et al., 2018).

The question asked for the highest educational qualification (for an illustration of the question see figure 5.1 or 5.2). The instruction below the question text said that respondents should give their last educational qualification, in case they were unsure about their "highest" qualification. Respondents with a foreign educational qualification had an additional instruction, which informed them to enter their response in the language in which they obtained the qualification.

5.4.2 Participants

We conducted the usability study at the GESIS Pretest Lab[4] in May 2015. 31 participants were recruited from the respondents pool maintained by the institute and through the use of advertisements. Before coming into the lab, each potential participant answered a series of screener questions over the phone to determine age, education, and country in which the highest educational qualification was obtained.

We based the selection of participants with German educational qualifications on age quotas (age categories 25–35 years, 36–50 years, 51–65 years) and education quotas (University degree versus one additional certificate after school graduation). We used these quotas to get a diverse picture from different educational levels, as Germany's educational system is very diverse because educational qualifications cover certificates from

[3]The underlying database of the combo box covered educational qualifications from 19 European countries and covered over 550 educational qualifications (see Schneider et al., 2018).

[4]The GESIS Pretest Lab is part of GESIS Leibniz Institute for the Social Sciences in Germany. This usability study was part of a question module for the Dutch LISS Panel 2016.

vocational training (Schneider and Kogan, 2008). Furthermore, we used age quotas to ensure that also elderly respondents could answer with the help of a computer and to cover out-dated qualifications. In each age group, we aimed for 2 participants with a university degree, whereas we aimed for 3 participants with another degree than a university degree (for an illustration see table 5.1) because research showed that other degrees than university degrees are more difficult to measure in Germany (Schneider et al., 2016).

Table 5.1: Quota of the German participants.

	Age			Overall
	25–35	36–50	51–65	
German educational qualifications				
University degree	2	2	2	6
Other degree (in addition to school)	3	3	3	9
Overall	5	5	5	15

As parts of the pretest aimed to investigate whether foreign educational qualifications could be measured with the combo box, we included participants with foreign educational qualifications. Participants with educational qualifications from abroad were selected without quotas. All 16 participants with foreign degrees had at least 12 years of schooling and were between 18 years and 61 years old.

Our final sample included 20 females and 11 males. 15 participants were born in Germany, and their mother tongue was German, whereas 16 participants were born in other European countries (for an overview see table A5.1 and for more details on the whole study see Lenzner et al., 2015). We ensured that all foreign-born participants spoke sufficient German to conduct the interview in German.

5.4.3 Procedure

The 31 respondents were seated in front of the eye tracker. After signing consent forms, the interviewers described the study and performed a calibration exercise with the eye tracker.

First, respondents answered an online questionnaire with at least six questions and a

maximum of 17 questions[5] on their education history that took about 10 minutes. During that time eye movements were recorded and interviewers[6] investigated the respondents' eye movements in real-time on a second computer screen in an adjacent room. Interviewers coded the eye movements on the spot and noted response difficulties.

Subsequently, interviewers conducted cognitive interviews (verbal probing). The probing questions for the interviews were pre-scripted and developed specifically to uncover why respondents used the combo box as they did and which response difficulties emerged. To remind respondents on how they answered the question and to talk in detail about the interface design respondents watched a video replay during the cognitive interviews. The video showed respondents' eye movements when navigating through the online questionnaire (see Neuert and Lenzner, 2016). The cognitive interviews took about 20 minutes.

After that, respondents participated in a usability study (with eye tracking and cognitive interviews) on the usage of search trees (nested response options) for another 20 minutes. Respondents received €30 as reimbursement for their participation.

5.5 Data

5.5.1 Eye-tracking data

We recorded eye movements with a Tobii T120 remote eye-tracking system (for further information see www.tobii.com), which has an eye-tracking camera within the frame of a 17" TFT monitor (resolution of 1280x1024 pixels). This eye-tracker permits unobtrusive eye movement recordings within head movements of 12x9x12 inches (30x22x30 cm). We used the corresponding Tobii Studio 3.3.1 software to analyze the data and to play back the eye movement recording, with (for post-hoc coding purposes) or without gaze data points (during the cognitive interviews).

We analyzed the eye-tracking data by predefining "areas of interest" (AOI). To quantify the gaze data of the respondents (120 gaze data points per second are collec-

[5]The number of questions differed between respondents depending on the number of educational qualifications mentioned.

[6]Six interviewers conducted at least three and a maximum of seven cognitive interviews (three researchers and three student assistants). All interviewers had received specific training regarding observing and coding eye movements.

ted), we defined three static and three dynamic AOIs (for an illustration see figure A5.1). The static AOIs covered the question text, the instruction, and the text field, whereas one dynamic AOI covered the combo box with response suggestions, and two AOIs divided the response list into two parts (top and bottom). In the following we use the total fixation time on a given AOI (as previously done by Höhne and Lenzner, 2017; Kamoen et al., 2017; Menold et al., 2014).

Eye-tracking is based on the assumption that eye movements match the visual attention (Rayner, 1977) and thus, eye-tracking assumes a relationship between fixations, gaze patterns and cognitive processing (Just and Carpenter, 1980; Staub and Rayner, 2007). If the assumption is true, then respondents with less cognitive processing should have fewer fixations; or shorter fixation times, while respondents with higher cognitive processing should have more fixations or longer fixation times.

5.5.2 Data from cognitive interviews

While respondents were filling out the survey, interviewers coded their eye movements. For this purpose, we pre-defined codes to identify how respondents used the combo box and to indicate difficulties in the question-answering process. Based on these codes interviewers asked probes during the subsequent cognitive interview.

The interviewers identified whether the respondents answered the combo box as a closed-ended question (using the drop-down box) or as an open-ended question (using the text field). The descriptions for assigning a code for the usage were:

- Respondent types something in the text field, selects a response from the drop-down box and presses the submit button.
- Respondent types something in the text field and presses the submit button without selecting a response from the drop-down box (response options were displayed).

Furthermore, interviewers coded difficulties in the question-answering process for each respondent based on the following coding scheme for eye movements:

- Respondent fixates longer on the question text or reads it repetitively.
- Respondent clicks on the submit button before typing something in the text field.

- Respondent gazes back and forth between question text, text field, and/or response options.
- Respondent types something in the text field, looks at the response options provided, deletes the entry and types something new.
- Respondent fixates longer on the displayed response options or repeatedly reads the response options.

In the subsequent cognitive interviews, interviewers asked probing questions based on the aforementioned coding scheme. For example, when respondents did not select one of the suggested response options, the interviewer asked: "While you were typing a list of response options appeared. Why haven't you selected any of these options?". Furthermore, interviewers asked questions on how difficult it was to answer the question with the help of the combo box, whether the response suggestions were inaccurate, whether they were astonished about the drop-down box, and when using the drop-down box whether they have chosen their intended answer (for details see table A5.1). Finally, interviewers asked respondents to explain the usage of the combo box to an imaginary friend.

5.6 Results

Table 5.2 reports the results of the eye-tracking data overall and separately for the two user types (closed-ended vs. open-ended). Table 5.2 shows the median fixation times, and the median fixation counts[7] of the different parts of the question for all respondents and the different usage types.

In table 5.3 we present the results of the cognitive coding separately for respondents who used the combo box closed-ended and for respondents who used the combo box open-ended. In the following, we describe different usage types by combining the data from the cognitive interviews and the eye-tracking data.

On average, respondents spend about the same amount of time looking at the question as they are looking at the instructions (3.07 vs. 3.41 seconds). Respondents' fixa-

[7] We used the medians because we detected outliers for which we did not want to adjust due to the small sample size.

Table 5.2: Median fixation times in seconds and median frequency of fixation counts on different parts of the combo box.

	Overall	n	Used the combo box closed-ended	Used the combo box open-ended
Question				
fixation time	3.07	30	2.48	3.18
fixation count	16		18	16
Instruction				
fixation time	3.41	30	4.20	3.03
fixation count	16		22	13
Text field				
fixation time	1.24	30	1.85	0.97
fixation count	5		7	4
Drop-down box				
Overall				
fixation time	2.63	21	2.69	2.03
fixation count	11		11	8
Look at the top				
fixation time	1.88	19	1.97	1.08
fixation count	7		7.5	5
Look at the bottom				
fixation time	1.25	16	1.42	1.02
fixation count	5		7	3
Look at it while typing				
fixation time	0.3	20	0.32	0.20
fixation count	1		2	1
Number of respondents			17	13

NOTE. – Due to problems with the recording of the eye fixation times respondent ID 28 got excluded. n = number of observations.

tion times on the empty text field is about one-third of the time they spend looking at the question or the instructions (1.21 seconds). Respondents with a foreign degree fixated the instructions longer and more often than respondents with a German degree (might be due to the shorter instructions for Germans).

When we examine the use of the combo box and more specifically the list of re-

Table 5.3: Coding of eye movements when respondents answered the combo box (multiple codes possible).

Respondent...	n	Respondent ID
Used the combo box closed-ended		
... types something in the text field and selects something from the response suggestions.	18	02, 05, 06, 07, 08, 11, 15, 17, 18, 19, 21, 22, 24, 26, 27, 29, 28, 31
... types something in the text field, looks at the response options provided, deletes the entry and types something new.	5	08, 21, 22, 24
... fixates long on the response options displayed or reads them repetitively.	3	11, 26
... gaze switches between question text, response options and/or text field.	1	19
Used the combo box open-ended		
... types something in the text field and presses the submit button without selecting a response from the list (options were displayed).	6	09, 12, 14, 16, 20, 30
... types something in the text field, looks at the response options provided, deletes the entry and types something new.	1	30
... fixates long on the response options displayed or reads them repetitively.	1	30
... types something in the text field without reading response suggestion (options were displayed).	7	01, 03, 04, 10, 13, 23, 25

NOTE. – Codes are based on real time and post-survey interviewer coding. Codes that did not occur are not listed. n = number of observations.

sponse suggestions below the text field we find that seven respondents did not see the provided list of response suggestions (the drop-down box). Respondents who saw the drop-down box fixated the response suggestions shorter than the question and the instructions (2.63 seconds). The short fixation times on the drop-down box support the fact that the vast majority of respondents felt that it was very easy or rather easy to give their response in the combo box (19 cases out of 22 cases; overall 27 cases out of 31 cases felt it was easy). Furthermore, we find that more respondents focused on the top of the list (19 cases out of 22 cases) than at the bottom of the list (16 cases out of 22 cases).

When we compare respondents, who used the combo box as a closed-ended question with respondents who used the combo box as an open-ended question we find differences in the median fixation times and the median fixation counts. Respondents who used the combo box with the help of the text field fixated the question text longer than the respondents who used the combo box with the help of the drop-down box. Furthermore, the median fixation time and count of the instructions was much lower for respondents who used the combo box in an open-ended format compared to respondents who used the combo box in a closed-ended format (4.20 seconds vs. 3.03 seconds; 22 fixations vs. 13 fixations). We also see that respondents using the text field in contrast to respondents who used the drop-down box have lower median fixation times and fixation counts on the drop-down box. Especially, the bottom of the drop-down box is fixated less often from respondents who used the text field.

To gain knowledge on how respondents understand the usage of the combo box, we asked them to imagine how they would explain the use of the combo box to a friend with low computer literacy. Almost all respondents explained all functions of the combo box (25 cases), for instance, that it was possible to either type in a response or to choose a response from the list. ID 16 explained the combo box as "[...] (s)he [one] should start typing, the first character or the first word and then (s)he should examine whether it [the educational qualification] is in their [in the list]. If the answer is yes, click on it [the educational qualification]; if the answer is no, continue typing [in the text field].". Nevertheless, most respondents who saw the drop-down box were astonished to get response suggestions (16 cases out of 22 cases). We asked respondents who saw the response suggestions whether their entered response reflected their intended response and the majority of respondents found their exact response (13 cases out of 18 cases). In the following, we elaborate on the use of the combo box open-ended or closed-ended question format.

Used the combo box closed-ended Overall, 18 respondents used the combo box closed-ended to answer the question (out of these respondents ten had a foreign degree). Respondents using the drop-down box fixate the response suggestions (on average 9.65 response suggestions were displayed) longer and more often compared to respondents who used the combo box text field.

Two respondents had relatively long fixation times on the response suggestions (re-

spondent with ID 11 = 65.81 seconds; respondent with ID 26 = 9.43 seconds). There were two reasons for this behavior. First, the respondents found it challenging to choose a response suggestion, because one option was a rather generic term, whereas the other was a very specific term (respondent with ID 26 decided for the generic term). Second, the respondent with ID 11 could not find his/her intended response (saw 20 response suggestions) and chose eventually not his/her highest educational qualification (from a researcher perspective).

One respondent had many gaze switches between the question, the text field and/or the answer options (respondent ID 19 had 18 fixations on the instructions and 11 fixations on the text field). When asking (s)he remembered that s(he) was surprised that response suggestions popped up and (s)he did not expect the response suggestions to be that accurate.

Five respondents started typing something into the text field and adjusted their intended response to the list of response suggestions presented in the drop-down box (respondent IDs: 08, 21, 22, 24, 30). Out of the five respondents with this response process, four respondents had a foreign degree (respondent IDs: 21, 22, 24, 30). Although the respondents with a foreign degree fixated the instructions on average longer and more often than the respondents with German degrees, two respondents read the instructions only briefly (respondent IDs: 22, 24). The cognitive interview revealed that the respondent with IDs 22 and 24 thought they had to translate their foreign degree into German. However, the moment these respondents saw the response suggestions they switched languages and gave their educational qualification in the language where the qualification was obtained (respondent IDs: 22, 24). Two other respondents did not find their intended response on the list and tried different search terms until they found the correct adaptation to their intended response (respondent IDs: 08, 30). Finally, one respondent (respondent with ID 21) deleted parts of his/her initial response and selected one of the response suggestions. It seems as if (s)he thinks one has to select a response within the list of response suggestions even though the initial answer was correct. When (s)he explained the functionality of the combo box during the cognitive interview, it turned out that (s)he did not know that there could have been multiple response suggestions for one educational qualification (due to synonyms).

Seven respondents stopped typing after several characters to investigate the list of response options (respondent IDs: 02, 07, 15, 19, 24, 26, 31). Those respondents were

asked to give their initial response. For one respondent the list of response suggestions caused confusion about the correct term of his/her educational qualification (respondent with ID 31). Three respondents indicated that they found their intended response in the list of response suggestions (respondent IDs: 07, 15,19). The other three respondents would have given other responses in a standard text field than with the combo box (respondents IDs: 02, 24, 26). These initial responses would have been either more specific or more generic than the response suggestions these respondents got (respondents IDs: 02, 24, 26).

In the cognitive interviews, we asked all respondents whether they found their intended response in the drop-down box. Five respondents indicated that they found an answer close to their intended answer (respondent IDs: 06, 22, 24, 27, 28) because they had a more specific term or a more contemporary term for their educational qualification in mind.

Although the majority of respondents felt that it was easy to give a response in the combo box (16 cases out of 18 cases), two respondents indicated that it was rather difficult for them to respond with the help of the combo box (respondent IDs: 11, 27). The respondent with ID 11 did not find his/her intended response and the respondent with ID 27 had to adjust his/her response to the response suggestions.

The majority of respondents who used the drop-down box of the combo box was surprised about the list of response suggestions (12 cases; respondent IDs: 02, 06, 07, 08, 15, 17, 19, 22, 24, 26, 28, 31). Based on the cognitive interviews we find two reasons why respondents were astonished. First, respondents expected a visual cue for an extension of the text field into a drop-down box. For example, the respondent with ID 06 explained his/her astonishment by saying, that [...] there was nothing on the right [corner of the text field]. Respondents with ID 31 and ID 06 were also missing an arrow at the right corner of the text field. Two respondents (respondent IDs: 11, 15) were uncertain whether it was possible to simply write something in the text field without selecting a response from the list of suggestions. For instance, the respondent with ID 11 explained the usage of the combo box as follows: "You do not see at the beginning whether it is a [standard] text field or whether there are predefined [educational] qualifications provided. Because of this, I would say try and type the first character and either you can continue writing, or you see a list of responses in which you find your

occupation or qualification." [8]

Second, respondents with foreign degrees were astonished that their response suggestions were in the language of their obtained educational qualification. In addition, the respondent with ID 17 says that s(he) would not have needed the suggestions, as (s)he had no problem remembering the term of his/her educational qualification and its spelling in the foreign language.

Used the combo box open-ended 13 respondents used the text field of the combo box to make a response (out of these six with a foreign degree). The average number of response suggestions the respondents saw was 3.80 response options.

Six respondents saw the list of response suggestions, but they did not select a response (respondent ID: 09, 12, 14, 16, 20, 30). We identified three reasons for this combo box usage with the help of the cognitive interviews. First, three respondents indicated that they did not find their intended response and hence, the list of response suggestions was not helpful (respondent IDs: 09, 16, 30), as the list did not include the intended answer. For example, the respondent with ID 16 said: "I have a special form of Abitur (similar to a university-entrance diploma), which you can only attain from one school in France.". The respondent with ID 30 commented on the presentation of the response suggestions: "I was irritated that the many answer options were not alphabetically or hierarchically structured. Some options were very specific others were very general, and again others were missing.". The respondent with ID 30 used the combo box with all its functions, for instance using different search terms, examining the response suggestions and finally using it as a standard text field. Although respondent with ID 30 looked at the response suggestions for 141.97 seconds s(he) found it easy to give a response in this format.

Second, the respondent with ID 14 indicated that it was less time consuming to type the response instead of selecting a response suggestion. The eye movements of this respondent indicated that (s)he looked briefly at the response suggestions and continued typing. This is in contrast to the other five respondents who saw the drop-down box but did not choose a response from it, who gazed at the text field about five times while typing.

[8] With this example, we also see that the constructs of education and occupation are easily being mixed in Germany. However, we do not want to go into detail on the construct validity.

Third, ID 12 and ID 20 did not know that they also could have clicked on a response suggestion. This is in line with respondents with ID 12 and ID 20 question-answering processes. The respondent with ID 12 observed the right answer but continues writing instead of selecting the answer. Although the respondent with ID 12 was not astonished about the response suggestion, not all functions of the combo box were well-defined for him/her.

In addition, two respondents were astonished that response suggestions popped up (respondent IDs: 14, 20). For example, the respondent with ID 20 said: "Normally there is a text or an arrow on which you can click and then the search starts, but there was nothing.".

Seven respondents (respondent IDs: 01, 03, 04, 10, 13, 23, 25) did not see the response suggestions at all. These respondents used the text field as a standard text field. When we examine the median fixation times of these specific respondents (respondent IDs: 01, 03, 04, 10, 13, 23, 25), we find that they fixate the question the longest (3.30 seconds), whereas they fixate the empty text field only briefly compared to the other respondents (0.88 seconds). Three respondents, who did not see the drop-down box, did not look at the screen while typing (respondent IDs: 03, 13, 25). The other four respondents looked at the screen once (respondent IDs: 01, 04, 10, 23); however, they did not see the response suggestions. The moment these respondents looked at the screen the drop-down box was already gone, as the entered response differed slightly from the response suggestions after a couple of letters.

Two respondents, who used the text field, did not know that they had the opportunity to click on a response suggestion instead they thought the combo box was a standard text field (respondent IDs: 03, 10). In the cognitive interview, they explicitly mentioned that they missed a visual cue which indicated possible response suggestions, such as an arrow.

5.7 Discussion

This article draws attention to respondents' use of combo boxes and its associated response difficulties. Based on the combination of eye-tracking data and cognitive interviews we investigate why respondents use the combo box either in a closed-ended ques-

tion format or in an open-ended question format. Furthermore, we examine whether using the combo box as a closed-ended or an open-ended question format relates to specific response difficulties.

We identified problems regarding the usage of the combo box and regarding the response suggestions itself. With regard to the usage of the combo box, we find that many respondents were not aware of all functions of the combo box. For example, it was not obvious to respondents that the combo box could have been used as a standard text field or as a drop-down box in the survey situation. This is in contrast to the cognitive interview situation where most respondents explained the functionality of the combo box correctly. A reason could be that respondents made assumptions about the usage when answering the questions in the cognitive interview. In addition, it was not obvious to respondents that response suggestions popped up below the combo box, as seven respondents did not see the response suggestions at all. These differences in usage demand different optimizations of the combo box.

To clarify the usage of the combo box, we recommend implementing visual cues, such as a magnifying lens in the left corner of the text field and/or an arrow in the right corner of the text field to indicate that respondents can submit a search term. These visual cues are commonly used in internet search engines. Actually, the respondent with ID 06 draw an analogy between the combo box and an internet search engine: "I would say, it is the same as Google, everybody knows it [Google]. One starts typing and with each character something pops up. And I have to type in letters until I get what I was intended to get." In the near future, these visual cues might become design convention and hence, their meaning is ingrained in respondents. Hence, survey practitioners should consider the usage of the combo box further, for example with a watermark in the text field itself, saying "Please type and/or click here" (this was also recommended by Couper and Zhang, 2016). Thus, visual and verbal cues could increase the understanding of the combo boxes' usage.

With regard to the response suggestion, respondents who used the drop-down box were irritated when both generic and specific response suggestions popped up. In this case, respondents were unsure which response they should have chosen, as response suggestions were neither "right" nor "wrong". Furthermore, at least one respondent felt overwhelmed by the number of response suggestions. In both cases, we suggest an adjustment of the search algorithm of the combo box. For example, the search algorithm

could be adjusted so that either the specific terms or the generic terms occur after a certain number of characters entered (depending on whether researchers aim for the generic or the specific term). In addition, the number of response suggestions presented could be reduced to a certain number by adapting the search algorithm. Hence, an adjustment of the search algorithm could facilitate respondents in choosing their answer from the list of response suggestions.

Finally, there were some specific problems with respondents with a foreign degree, which are only relevant for cross-national surveys or surveys with migrants. First, they initially did not answer the question in the language of their educational qualification obtained. To clarify this, a watermark saying "Please type and/or click here" could be displayed in the language of origin to making it clear which language the researcher expects from the respondent. Furthermore, some respondents with foreign degree indicated that they could not find their intended answer in the list of response suggestions. This will be resolved with the time, as the database will be updated with unknown educational qualifications after it has been checked whether the educational qualification is an officially recognized term.

5.8 Conclusions

So far long-list questions have been either asked in an open-ended or a closed-ended question format. Technological advancements allow the usage of combo boxes in computer-assisted interviews which combine open-ended and closed-ended questions. Although, combo boxes are widely used we do not know how respondents use combo boxes and whether combo boxes have similar problems as open-ended and closed-ended question formats. We presented an eye-tracking study in combination with cognitive interviews, to explore the usability of combo boxes.

Our qualitative study has two major findings: first, respondents use combo boxes differently and hence, they perceive the visual cues of combo boxes differently. We find that slightly more respondents use the combo box like a drop-down box. However, these respondents were unsure whether the combo box could be used as a standard text field which means respondents think they have to map their intended response to one of the response options given. On the one hand, knowing that respondents would not enter

their intended answer in the text field might relativize the usage of combo boxes instead of drop-down boxes (do not use the text field as an "Other" category). On the other hand, combo boxes initially do not display response options which have the advantage that respondents are less influenced in their initial memory retrieval and judgment by the responses which were defined by the researchers.

Second, we find problems which are normally associated with open-ended and closed-ended questions. For example, five out of 18 respondents who used the combo box like a drop-down box did not find their intended answers, whereas four out of 13 respondents who used the combo box as a standard text field did not give the researchers' intended answers. Nevertheless, the majority of respondents gave their intended and the researchers' intended answer.

Unfortunately, the sample size in this study is relatively small (n = 31), which does not allow for a generalization of the results to the general population. Furthermore, the group of respondents was heterogeneous in terms of country of origin, which resulted in additional response difficulties (e.g., respondents were not aware of answering in their language of origin). However, the heterogeneity of the group has the positive side effect that we simultaneously tested whether combo boxes can be used in cross-national contexts and which difficulties arise with this subgroups of respondents.

Future work would benefit from a comparison of different designs of the combo box against other interface designs of long-list questions, such as search trees or drop-down boxes. Similarly, the recommendations for visual cues regarding the interface design of combo boxes need to be examined further. Specifically, the adjustment of the search algorithm in terms of generic and specific response suggestions needs further testing, as respondents with different educational qualifications may experience this problem differently.

In addition, an adaptation of the interface design to smartphones and tablets needs further usability testing because the usage requirements differ from those of computers. Furthermore, the combo box was based on JavaScript which can cause trouble when respondents switch off JavaScript in their browsers or have an old version of JavaScript. However, survey practitioners sometimes check in online panels whether respondents can use interface designs which are programmed with JavaScript (for example the LISS Panel in the Netherlands).

Despite the limitations in the sample composition, our findings provide encourage-

ment for survey designers who want to support respondents in their question-answering process when asking long-list questions. Furthermore, the implementation of the combo box in cross-national surveys, which include various groups of migrants, has the advantage that almost all educational qualifications can be covered. Moreover, the utility of these findings for survey organizations lies in its transferability to other questions with long lists of answer options, such as brand names, occupations or prescription drugs.

Due to technical advancements survey researcher can develop alternative interface designs for long list questions. Although, combo boxes are already implemented in the survey, our results indicate that the use of combo boxes is not yet so ingrained that respondents do not have to think about the functionality of a combo box. Therefore, combo boxes need visual and verbal cues to explain its functionality. This work and its recommendations for optimizing the design of combo boxes represent one step in tapping the full potential of combo boxes.

Acknowledgments

This research was supported by grant SAW-2013-GESIS-5 249 provided by the Leibniz-Association through the Leibniz competition. We thank Silke Schneider, Franziska Adams, Roberto Briceno-Rosas, Katharina Disch, Stefanie Gebhardt, Uta Landrock, Timo Lenzner, Maurice Martens, Verena Ortmanns, and Wanda Otto for their support.

References

Callegaro, M., Manfreda, K. L., and Vehovar, V. (2015). *Web Survey Methodology*. Sage, London, UK.

Couper, M. P. (2000). Usability Evaluation of Computer-Assisted Survey Instruments. *Social Science Computer Review*, 18(4):384–396.

Couper, M. P., Kennedy, C., Conrad, F. G., and Tourangeau, R. (2011). Designing Input Fields for Non-Narrative Open-Ended Responses in Web Surveys. *Journal of Official Statistics*, 27(1):65–85.

Couper, M. P., Tourangeau, R., Conrad, F. G., and Crawford, S. D. (2004). What They See Is What We Get: Response Options for Web Surveys. *Social Science Computer Review*, 22(1):111–127.

Couper, M. P. and Zhang, C. (2016). Helping Respondents Provide Good Answers in Web Surveys. *Survey Research Methods*, 10(1):49–64.

Foddy, W. (1993). *Constructing Questions for Interviews and Questionnaires: Theory and Practice in Social Research*. Cambridge University Press.

Fowler, F. J. (1995). *Improving Survey Questions: Design and Evaluation*, volume 38 of *Applied Social Research Methods Series*. Sage Publications, Thousand Oaks, CA.

Höhne, J. K. and Lenzner, T. (2017). New Insights on the Cognitive Processing of Agree/Disagree and Item-Specific Questions. *Journal of Survey Statistics and Methodology*.

Jenkins, C. R. and Dillman, D. A. (1997). Towards a Theory of Self-Administered Questionnaire Design. In Lyberg, L. E., Collins, M., de Leeuw, E. D., Dippo, C., Schwarz, N., and Trewin, D., editors, *Survey Measurement and Process Quality*, Wiley Series in Probability and Statistics, chapter 7, pages 165–196. John Wiley & Sons, New York, NY.

Just, M. A. and Carpenter, P. A. (1980). A Theory of Reading: From Eye Fixations to Comprehension. *Psychological Review*, 87(4):329.

Kamoen, N., Holleman, B., Mak, P., Sanders, T., and van den Bergh, H. (2017). Why Are Negative Questions Difficult to Answer? On the Processing of Linguistic Contrasts in Surveys. *Public Opinion Quarterly*, 81(3):613–635.

Krosnick, J. A. and Alwin, D. F. (1987). An Evaluation of a Cognitive Theory of Response-Order Effects in Survey Measurement. *Public Opinion Quarterly*, 51(2):201–219.

Krosnick, J. A. and Presser, S. (2009). Question and Questionnaire Design. In Wright, J. D. and Marsden, P. V., editors, *Handbook of Survey Research*, chapter 9, pages 263–315. Elsevier, San Diego, CA.

Lenzner, T., Kaczmirek, L., and Lenzner, A. (2010). Cognitive Burden of Survey Questions and Response Times: A Psycholinguistic Experiment. *Applied Cognitive Psychology*, 24(7):1003–1020.

Lenzner, T., Neuert, C. E., Otto, W., Landrock, U., Adams, F., Disch, K., Gebhardt, S., and Menold, N. (2015). *CAMCES - Computer-Assisted Measurement and Coding of Educational Qualifications in Surveys. Kognitiver Pretest*. GESIS Projektbericht. Version: 1.0. GESIS - Pretestlabor.

Menold, N., Kaczmirek, L., Lenzner, T., and Neusar, A. (2014). How Do Respondents Attend to Verbal Labels in Rating Scales? *Field Methods*, 26(1):21–39.

Neuert, C. E. and Lenzner, T. (2016). A Comparison of Two Cognitive Pretesting Techniques Supported by Eye Tracking. *Social Science Computer Review*, 34(5):582–596.

Presser, S., Couper, M. P., Lessler, J. T., Martin, E., Martin, J., Rothgeb, J. M., and Singer, E. (2004). Methods for Testing and Evaluating Survey Questions. *Public Opinion Quarterly*, 68(1):109–130.

Rayner, K. (1977). Visual Attention in Reading: Eye Movements Reflect Cognitive Processes. *Memory & Cognition*, 5(4):443–448.

Redline, C. D., Tourangeau, R., Couper, M. P., Conrad, F. G., and Ye, C. (2009). *The Effects of Grouping Response Options in Factual Questions with Many Options*. Annual Conference of the Federal Committee on Statistical Methodology, Washington, DC.

Reips, U.-D. (2000). The Web Experiment Method: Advantages, Disadvantages, and Solutions. In Birnbaum, M. H., editor, *Psychological Experiments on the Internet*, chapter 4, pages 89–117. Elsevier, San Diego, CA.

Reja, U., Manfreda, K. L., Hlebec, V., and Vehovar, V. (2003). Open-Ended vs. Close-Ended Questions in Web Questionnaires. *Developments in Applied Statistics*, 19(1):159–177.

Schneider, S. L., Briceno-Rosas, R., Herzing, J. M. E., and Ortmanns, V. (2016). Overcoming the Shortcomings of Long List Showcards: Measuring Education With an Adaptive Database Lookup. In *9th International Conference on Social Science Methodology*, RC33 Conference, Leicester, UK.

Schneider, S. L., Briceno-Rosas, R., Ortmanns, V., and Herzing, J. M. E. (forthcoming 2018). Measuring Migrants' Educational Attainment: The CAMCES Tool in the IAB-SOEP Migration Sample. In Behr, D., editor, *Surveying the Migrant Population: Consideration of Linguistic and Cultural Aspects*. GESIS Schriftreihe, Cologne, DE.

Schneider, S. L. and Kogan, I. (2008). The International Standard Classification of Education 1997: Challenges in the Application to National Data and the Implementation in Cross-National Surveys. In Schneider, S. L., editor, *The International Standard Classification of Education (ISCED-97). An Evaluation of Content and Criterion Validity for 15 European Countries*, chapter 1, pages 13–46. MZES, Mannheim, DE.

Schwarz, N. (1999). Self-Reports: How the Questions Shape the Answers. *American Psychologist*, 54(2):93–105.

Schwarz, N. and Sudman, S. (1996). *Answering Questions: Methodology for Determining Cognitive and Communicative Processes in Survey Research*. Wiley: Jossey-Bass, Hoboken, NJ.

Smyth, J. D., Dillman, D. A., Christian, L. M., and McBride, M. (2009). Open-Ended Questions in Web Surveys: Can Increasing the Size of Answer Boxes and Providing Extra Verbal Instructions Improve Response Quality? *Public Opinion Quarterly*, 73(2):325–337.

Smyth, J. D., Dillman, D. A., Christian, L. M., and Stern, M. J. (2006). Effects of Using Visual Design Principles to Group Response Options in Web Surveys. *International Journal of Internet Science*, 1(1):6–16.

Staub, A. and Rayner, K. (2007). Eye Movements and On–line Comprehension Processes. In Gaskell, M. G., editor, *The Oxford Handbook of Psycholinguistics*, chapter 19, pages 327–342. Oxford University Press, Oxford, UK.

Tijdens, K. (2015). Self-Identification of Occupation in Web Surveys: Requirements for Search Trees and Look-up Tables. *Survey Methods: Insights from the Field*. https://doi.org/10.13094/SMIF-2015-00008.

Tourangeau, R., Conrad, F. G., and Couper, M. P. (2013). *The Science of Web Surveys*. Oxford University Press, New York, NY.

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

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

Appendix

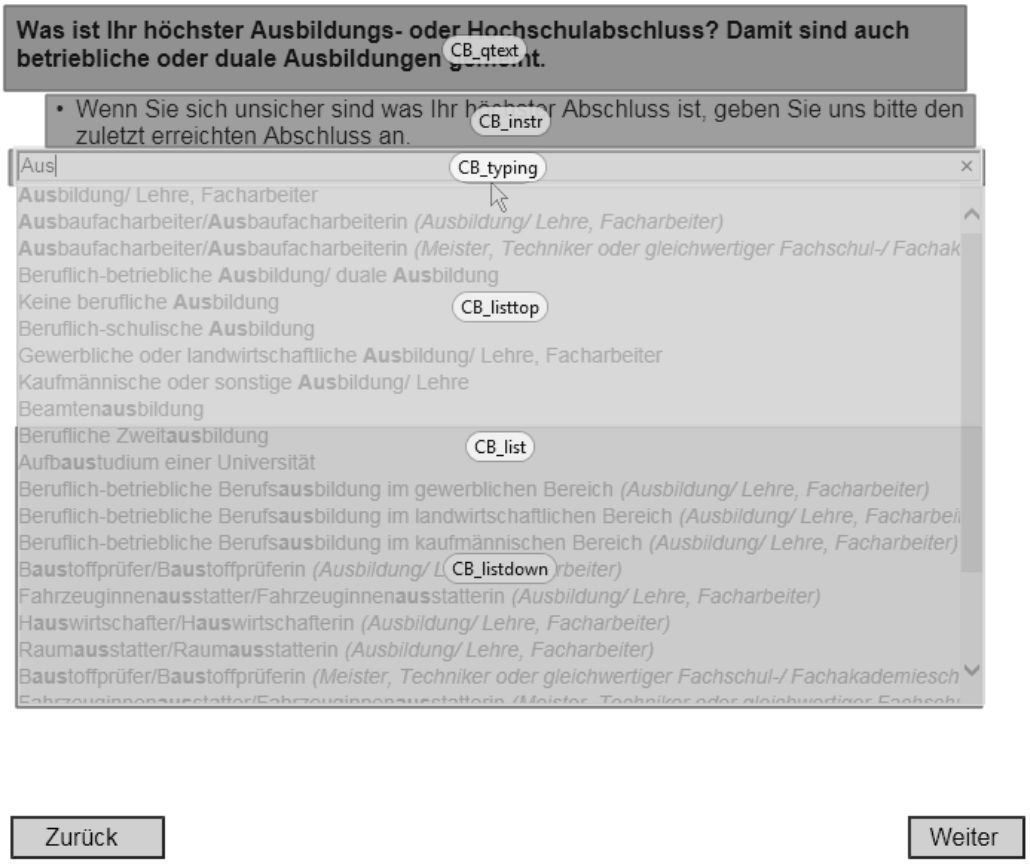


Figure A5.1: AOIs for the interface design of the combo box after three characters were typed in.

Table A5.1: Summary of respondents' characteristics, answers to questions and codes of the cognitive interviews.

Respondent ID	Age group	Country born	Highest educational qualification	Used the combo box	Recorded response difficulties	Difficult	Inaccurate	Astonished	Explained incorrectly	Respondents' intended answer	Researchers' intended answer
01	25-35	German	university degree	open-ended	yes	no	no	not applicable	no	not applicable	no
02	25-35	German	university degree	closed-ended	no	no	no	yes	no	yes	yes
03	36-50	German	university degree	open-ended	yes	no	no	not applicable	yes	not applicable	yes
04	36-50	German	university degree	open-ended	yes	no	no	not applicable	no	not applicable	yes
05	51-65	German	university degree	closed-ended	no	no	no	no	no	yes	yes
06	51-66	German	university degree	closed-ended	no	no	yes	yes	no	yes	yes
07	25-35	German	no university degree	closed-ended	no	no	no	yes	no	yes	yes
08	25-35	German	no university degree	closed-ended	yes	no	no	yes	no	yes	yes
09	25-35	German	no university degree	open-ended	yes	no	no	no	no	not applicable	yes
10	36-50	German	no university degree	open-ended	yes	no	no	not applicable	yes	not applicable	no
11	36-50	German	university degree	closed-ended	yes	yes	no	no	yes	no	yes
12	36-50	German	no university degree	open-ended	no	no	no	no	yes	yes	yes
13	51-65	German	no university degree	open-ended	yes	no	no	not applicable	no	not applicable	yes
14	51-65	German	no university degree	open-ended	yes	no	no	yes	no	yes	yes
15	51-36	German	university degree	closed-ended	no	no	no	yes	yes	yes	yes
16	18-24	Foreign	no university degree	open-ended	yes	no	no	no	no	not applicable	yes
17	25-35	Foreign	university degree	closed-ended	no	no	no	yes	no	yes	yes
18	36-50	Foreign	university degree	closed-ended	no	no	no	no	no	yes	yes
19	51-65	German	university degree	closed-ended	no	no	no	yes	no	yes	yes
20	18-24	Foreign	no university degree	open-ended	yes	no	no	yes	yes	yes	no
21	18-24	Foreign	no university degree	closed-ended	yes	no	no	yes	no	yes	yes
22	25-35	Foreign	university degree	closed-ended	yes	no	yes	yes	no	almost	yes
23	36-50	Foreign	university degree	open-ended	yes	no	no	not applicable	no	not applicable	yes
24	36-50	Foreign	university degree	closed-ended	yes	no	yes	yes	no	almost	yes
25	18-24	Foreign	no university degree	open-ended	yes	no	no	not applicable	no	not applicable	yes
26	18-24	Foreign	no university degree	closed-ended	yes	no	no	yes	no	yes	yes
27	18-24	Foreign	no university degree	closed-ended	no	yes	yes	no	no	almost	yes
28	51-65	Foreign	no university degree	closed-ended	no data	no	yes	yes	no	almost	no
29	18-24	Foreign	no university degree	closed-ended	no	no	no	not applicable	no	yes	yes
30	36-50	Foreign	university degree	open-ended	yes	yes	no	no	no	no	no
31	36-50	Foreign	university degree	closed-ended	no	no	no	yes	no	yes	yes

Chapter 6

Conclusions

6.1 Coping with technological and societal change

From a technological perspective one may argue that the digital revolution is over; however, from a societal perspective, the digital revolution is in full swing. This gap between technological advancements and human adaptation to technology is challenging survey research. On the one hand, survey practitioners are adapting survey designs to technological innovations, while potential respondents have yet to adapt to such changes. Thus, survey practitioners need to be hesitant with the incorporation of new technologies into the survey design, as there is the chance of overstraining respondents' abilities. Consequently, it should be the aim of survey researchers to create data collection methods/instruments that are neither too far ahead nor behind their targeted respondents.

Over the course of technological change ,computer-assisted surveys and especially online surveys have become a prevalent data source for market research and behavioral sciences (Baker et al., 2010, p. 7; Schonlau et al., 2002). Online surveys are attractive because they enable survey practitioners to conduct interviews cost-effectively in terms of time, space, and labor (Greenlaw and Brown-Welty, 2009; Hardigan et al., 2012; Kaplowitz et al., 2004). However, there are concerns about the generalizability of estimates to the general population based on online surveys (Best et al., 2001; Bethlehem, 2010; Dever et al., 2008; Mohorko et al., 2013; Sterrett et al., 2017). To conduct high quality surveys probability-based online panels emerged, which acknowledge the need for both probability sampling and coverage of persons without computers and/or internet (see for example Blom et al., 2017; Bosnjak et al., 2013; de Vos, 2010; Knoef and de Vos, 2009; Revilla et al., 2016). For this purpose, probability-based online panels either equip non-internet households with devices and internet, or non-internet households are interviewed via a different mode, for example with mail questionnaires (for examples see Blom et al., 2016; Bosnjak et al., 2017). Nevertheless, about one fourth of those who use the internet for private purposes prefer a paper questionnaire over an online questionnaire in mixed-mode surveys (Pforr and Dannwolf, 2017) and even when offering non-internet households the necessary devices to participate in an online survey these households are less likely to do so (Blom et al., 2017). Hence, the technological

change towards online surveys might not suit all targeted respondents. Consequently, we need further investigations into whether respondents' adaptation to technology is a reason for peoples' hesitation to participate in online surveys.

Another development that comes along with technical advancements are alternative survey instruments in computer-assisted surveys. About ten years ago Couper (2008, p. 113) published a book on the design of online surveys in which he states that even when technical issues are overcome questions remain about if and when alternative interface designs are appropriate for survey measurement. Survey practitioners have to be aware of potential response differences and differences in respondents' knowledge and familiarity with alternative interface designs. Therefore, it is important to examine whether alternative interface designs increase respondents' effort, reduce attentiveness to the answer options, and influence the measurement with dependence on respondents' adaptation to technological change.

With this dissertation, I aim to fill the gap and contribute to a deeper understanding of how technological change influence measurement and unit nonresponse in computer-assisted surveys. For this purpose, I explored the association between survey design and different respondent groups. Paper I and II mainly focused on the effects of unit nonresponse in the recruitment process of a probability-based online panel, whereas paper III and VI addressed measurement quality of alternative interface designs for long list questions.

6.2 Summaries

Paper I In the first paper (chapter 2) we focused on the aspect of interviewer effects as an influence on nonresponse rates. To investigate interviewer effects on nonresponse, we use an alternative parametrization namely separate coding in random slopes of a multilevel model. On the one hand, we find small interviewer effect sizes for offliners, which indicates that for offliners it is tricky to increase the response propensities with interviewer-related fieldwork strategies (e.g., targeted response). On the other hand, we find large interviewer effect sizes for onliners suggesting that response propensities could be raised with interviewer-related fieldwork strategies. Ultimately, our proposed parametrization strategy for random slopes is important for survey practice, as it can be

used within fieldwork monitoring to inform effective recruitment strategies.

Paths for future research might include the search for tests of deviance that are applicable to a setting where separate coding is used for random slopes, as the standard tests of deviance are misleading for multilevel models, due to the χ^2 skewed distribution of variances estimations (Hox, 2010, pp. 45–47). We considered the solution of Molenberghs and Verbeke (2007) to be the best. Nevertheless, there is further research needed on this topic. Furthermore, from a frequentist paradigm, samples on each level of the multilevel model should be randomly selected. However, in survey research interviewers - the higher level - are not randomly selected. Although the frequentist view may be true, we relaxed this assumption as we argue that interviewers are a finite population and that we use it as a grouping variable in the multilevel model. In addition, we do not make causal inference about the interviewer effects on nonresponse; hence the nonrandom selection of interviewers is of minor concern in paper I.

Paper II In paper II (chapter 3) we investigate nonresponse mechanisms in a probability-based online panel. With a multi-dimensional classification of IT literacy, we predict response to the first online panel wave and participation across panel waves. Our results show that IT literacy is associated with both survey variables and sample units' propensity to respond. Hence, IT literacy is a predictor of nonresponse bias. Furthermore, we find that the different groups of IT literacy classes are associated with different attrition patterns. The importance of a multi-dimensional classification of IT literacy lies in its application during fieldwork monitoring and for nonresponse and attrition adjustments. Survey conductors can use respondents' IT literacy either during fieldwork monitoring to implement interviewer-related fieldwork strategies and reminder procedures or for later post-survey adjustments such as imputations or weighting.

Future research should explore the impact of using the four original indicators of the LCA to investigate whether the data reduction of the LCA has an impact on the results. Nevertheless, we used a LCA because the four observed variables are statistically dependent. The important aspect of LCA is that within each of the four latent classes, the four observed variables are statistically independent (see McCutcheon, 2002). As a result, the IT literacy class a respondent has, "causes" the IT usage behavior of a respondent. Consequently, it is the question of whether one is interested in the association of each individual variable with unit nonresponse or not. If the answer is yes, then one

has to address the question of how to include multiple variables in a model when they are related to each other. After all, we argued that it is more a combination of different variables namely IT literacy which causes unit nonresponse in the online panel. Therefore, the LCA was our first choice. In addition, future research should include other variables for the LCA, as we had only a limited number of variables available.

In addition, paper I and II used data from the GIP recruitment interview. Future research should investigate the influence of biases in the initial panel recruitment. Questions might include whether people with higher political interest are overrepresented in the initial GIP sample. However, we could not test this, as we do not have benchmarks of the population for the selected variables. Furthermore, our argument was based on the transition from the face-to-face interview to the online survey and hence, we did not investigate bias in the initial recruitment and whether the potential onset bias became more aggravated with online panel participation in paper II. Moreover, the problem of onset bias is negligible in paper I, as we were interested in interviewer effects and hence, do not draw inferences about population quantities.

Paper III In paper III (chapter 4) I scrutinized three alternative interface designs for long-list questions by using the example of highest educational qualification in Germany. I compared a long list with radio buttons, a combo box and a search tree with each other regarding response burden and data quality. My results indicate that combo boxes lower response burden. However, it turns out that the combo box comes along with much more post-survey coding than the long list or the search tree. There is a trade-off between a decrease in response burden and an increase of post-survey coding when implementing one of the alternative interface designs. These findings add to research on the optimal design for long-list questions in computer-assisted surveys and how we can facilitate the question-answering process without decreasing data quality by using technical innovations.

Although I reinterviewed the respondents and thus had the highest educational qualification from two-time-points, I cannot say whether the implemented interface designs are good in measuring the highest educational qualification. To evaluate which interface design works best I would have needed a "gold standard" measure of the highest educational qualification for the SOEP-IS respondents (e.g., by linking the data with administrative data). Therefore, I cannot say which interface design produces higher

validity. Furthermore, I cannot speak of reliability, as the interface designs were not exactly the same in 2011 and 2014. The issues of validity and reliability could be addressed in future research by linking the data to administrative data or running the same interface at two-time-points.

In addition, future research would gain from extending the experimental conditions by two conditions: open-ended question (corresponding the combo box) and multi-page filtering (corresponding the search tree). However, the sample size was too small for more experimental conditions (for information on power analysis see Döring and Bortz, 1995; Lachin, 1981). Furthermore, research has shown that open-ended education questions are error-prone (Herzing and Schneider, 2014) and multi-page filtering results in many dropouts in occupation questions (Tijdens, 2014). In addition, future research would gain from further insights about respondents' IT literacy. Information of respondents IT literacy could have added to the explanation of response burden between interface designs.

Paper IV In paper IV (chapter 5) we conducted a usability study with the aim to identify response difficulties when answering questions with the help of the combo box. Based on both eye-tracking and cognitive interviews we were able to compare respondents who used the combo box in an open-ended and a closed-ended question format. We find some response difficulties which are normally associated with open-ended and closed-ended questions. Furthermore, we detected specific response difficulties with regard to the functionality of the combo box. To increase the usability and decrease response difficulties when answering questions with a combo box we suggest to implement visual cues, such as a magnifying lens in the left corner of the text field to indicate a possible search. Moreover, we discover that the search algorithm could be optimized, to improve respondents' searches through the list of suggestions. Our findings are transferable to other questions where combo boxes are used. The understanding of the question-answering process given a combo box adds to general research on designing interfaces to aid respondents in answering long-list questions.

The eye-tracking study in paper IV would have benefited from a larger sample size (n = 31). A larger sample size in the eye-tracking study would allow statistical testing (for example see Galesic et al., 2008). However, Sudman (1983, p. 226) maintained that 20-50 cases are usually enough to find major difficulties in a question in cognitive

interviews. Furthermore, the sample of the usability study was a very heterogeneous group, as migrants were included. On the one hand, additional usability problems might be introduced, because respondents with a migration background read the question in German and had to answer in their language of origin. On the other hand, the inclusion of migrants in the sample had the advantage that we gained some knowledge on how the combo box is used by migrants in case survey practitioners implement the combo box in a cross-national survey. In this regard, future research could investigate whether the usability differs across cultures.

This dissertation used the survey lifecycle by Groves et al. (2011) as its framework. In paper I and II I refer to the quality perspective and more specifically to unit nonresponse. However, I do not quantify nonresponse bias in both papers. The design perspective served well in paper III and IV when we investigated the cognitive and the usability standards. Nevertheless, it was not the goal of this dissertation to test the underlying theoretical mechanism of measurement quality or unit nonresponse.

6.3 Conclusions

Tourangeau (2017, p. 811) stated in his presidential speech at the conference of the American Association of Public Opinion (AAPOR) [that] [w]e need to redouble our efforts to understand why people [do not] want to do surveys anymore. The issue why people do not participate in online surveys is raised in paper I and II. In paper I we find that interviewers are not the reason for the low response rates of offliners. Therefore, we have to find other reasons why offliners do not participate, and an apparent explanation is the mode. It is cognitively much more burdensome for offliners to participate in an online panel even though they get the means to participate. Hence, we have to think whether a mixed-mode approach would increase the response propensities of offliners.

In paper II we detect that IT literacy is a predictor of nonresponse and attrition in an online panel. With regard to the findings of paper II we could also think of using a mixed-mode approach to increase response rates among hesitant onliners. Furthermore, we know from paper I that interviewers affect nonresponse in case of onliners. Hence, interviewer-related fieldwork strategies which target reluctant IT literacy classes could be implemented. Moreover, IT literacy could be used to reduce potential non-

response bias by modeling daily response propensities across IT literacy classes and adjust reminder procedures specifically for specific IT literacy classes (for example see Vandenplas and Loosveldt, 2017).

It is yet not clear whether we can fix unit nonresponse bias in the data collection or whether we have to fix unit nonresponse bias with post-survey adjustments. Thus, the classes of IT literacy should be used for post-survey adjustments. The adjusted estimates could be compared to population values, and then we would know how good IT literacy works as an auxiliary variable for nonresponse bias.

In paper III and IV we investigated alternative interface designs for long-list questions. Nowadays, survey practitioners and potential respondents use multiple survey modes with which they conduct online surveys. Thus, the proposed interface designs in paper III and IV need to be adapted to tablets and smartphones. In addition, one could think of extending the combo box with voice input to make the usage with smartphones easier. Furthermore, survey research faces the challenge that surveys are increasingly relying on multiple devices. In order to achieve measurement equivalence between different devices, we have to develop questionnaire designs that are optimized for multiple modes. Therefore, combo boxes and search trees need further testing with different devices and regarding measurement equivalence.

6.4 Thoughts about future technological advancements in survey research

This dissertation covered technical advancements in survey research from two perspectives. First, whether respondents' adaptation to technology influences unit nonresponse. Second, whether alternative interface designs for long-list questions facilitate respondents in their question-answering process. Building on the studies of this dissertation I would like to tackle the issue of unit nonresponse and measurement in my future research.

Currently, survey practitioners still have the problem that they do not know whether they cover the target population when they use online surveys. Although the internet penetration rates are increasing and could be negligible in a couple of years, survey organizations will still face the problem of defining an appropriate sample frame for a

population. Yet, there has not been a link found to connect postal addresses with email addresses. In the future, the alternative sample frames might be developed where either a mix of devices is used, or neither email addresses nor postal addresses are needed.

While human adaptation to technology will improve in the near future, survey practitioners will still face the problem of decreasing response rates. Hence, we have to think about interactive fieldwork to reduce nonresponse and especially nonresponse bias in the first place. In the future, we could think of the integration of big data and fieldwork monitoring to better place reminders and survey requests. For example, based on the data gathered from smartwatches or smartphones estimations could predict whether a person is at home and has time to fill out an online survey or not. Consequently, sensor data could make fieldwork more efficient.

Furthermore, we have to think about conducting surveys on new devices such as smartwatches. From my point of view, it is conceivable that smartwatches display the questions and respondents' answers via voice input. Therefore, a promising branch of research is the development of surveys where respondents can answer via voice input.

With the development of new devices, multi-mode designs are becoming increasingly prevalent. Not only does unified questionnaire construction with a maximized design for multiple devices become a necessity, but researchers must consider the conditions under which respondents answer the survey questions. As most devices can be used mobile, we can assume that respondents fill out the questionnaire in everyday situations, for example, while sitting in the bus. The attention span of respondents will be shorter resulting in new challenges for survey researchers. Therefore, we also have to consider a responsive survey design to make questionnaires shorter (Brick and Tourangeau, 2017). In this context, big data might be used as additional information for the systematic missing data which is produced in responsive survey designs.

Everything will become intelligent; soon we will not only have smartphones, and smartwatches but also smart homes, smart factories, and smart cities. It is estimated that in 10 years there will be 150 billion networked measuring sensors, 20 times more than there are people on earth (Helbing et al., 2017). The amount of data we produce will further increase, and this data will contain information on how we think and feel. Currently, there is still a gap between the amount of data we generate and the possibility to process these amounts of data. But in the near future, we will be able to automatically analyze data in real time. However, we do not know whether we will get better predic-

tions from our "digital crystal ball" (this expression was taken from Helbing, 2016). These technical advancements will have an impact on survey research, as they can reduce survey costs and the response burden (i.e., by reducing the number of questions).

From my point of view the biggest challenges survey research faces in the future are the following: what to do with this amount of data, do we get better predictions from the "digital crystal ball", and how do we adjust survey designs to new technological advancements such as mixed reality lenses (e.g., Microsofts® HoloLens). Obviously, the survey world will continue to change, we just do not know how fast it will change. Nevertheless, the responsibility of survey researchers remains the same, to create high-quality data collection methods that are neither too far ahead nor behind our respondents.

References

Baker, R., Blumberg, S. J., Brick, J. M., Couper, M. P., Courtright, M., Dennis, J. M., Dillman, D., Frankel, M. R., Garland, P., Groves, R. M., Kennedy, C., Krosnick, J., Lavrakas, P. J., Lee, S., Link, M., Piekarski, L., Rao, K., Thomas, R. K., and Zahs, D. (2010). Research Synthesis: AAPOR Report on Online Panels. *Public Opinion Quarterly*, 74(4):711–781.

Best, S. J., Krueger, B., Hubbard, C., and Smith, A. (2001). An Assessment of the Generalizability of Internet Surveys. *Social Science Computer Review*, 19(2):131–145.

Bethlehem, J. (2010). Selection Bias in Web Surveys. *International Statistical Review*, 78(2):161–188.

Blom, A. G., Bosnjak, M., Cornilleau, A., Cousteaux, A.-S., Das, M., Douhou, S., and Krieger, U. (2016). 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. (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, M., Dannwolf, T., Enderle, T., Schaurer, I., Struminskaya, B., Tanner, A., and 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*. https://doi.org/10.1177/0894439317697949.

Bosnjak, M., Haas, I., Galesic, M., Kaczmirek, L., Bandilla, W., and Couper, M. P. (2013). Sample Composition Discrepancies in Different Stages of a Probability-Based Online Panel. *Field Methods*, 25(4):339–360.

Brick, J. M. and Tourangeau, R. (2017). Responsive Survey Designs for Reducing Nonresponse Bias. *Journal of Official Statistics*, 33(3):735–752.

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

de Vos, K. (2010). Representativeness of the LISS-Panel 2008, 2009, 2010.

Dever, J. A., Rafferty, A., and Valliant, R. (2008). Internet Surveys: Can Statistical Adjustments Eliminate Coverage Bias? *Survey Research Methods*, 2(2):47–60.

Döring, N. and Bortz, J. (1995). *Forschungsmethoden und Evaluation für Sozialwissenschaftler*. Springer, Berlin, DE.

Galesic, M., Tourangeau, R., Couper, M. P., and Conrad, F. G. (2008). Eye-tracking Data: New Insights on Response Order Effects and Other Cognitive Shortcuts in Survey Responding. *Public Opinion Quarterly*, 72(5):892–913.

Greenlaw, C. and Brown-Welty, S. (2009). A Comparison of Web-Based and Paper-Based Survey Methods Testing Assumptions of Survey Mode and Response Cost. *Evaluation Review*, 33(5):464–480.

Groves, R. M., Fowler, F. J., Couper, M. P., Lepkowski, J. M., Singer, E., and Tourangeau, R. (2011). *Survey Methodology*, volume 561 of *Wiley Series in Survey Methodology*. John Wiley & Sons, Hoboken, NJ, 2nd edition.

Hardigan, P. C., Succar, C. T., and Fleisher, J. M. (2012). An Analysis of Response Rate and Economic Costs between Mail and Web-Based Surveys among Practicing Dentists: A Randomized Trial. *Journal of Community Health*, 37(2):383–394.

Helbing, D. (2016). Ferngesteuert oder Selbstgesteuert–Perspektiven der Digitalen Gesellschaft (Remote or Self-Controlled–Perspectives of the Digitial Society). In *Presented at Conference Futures 2015*, Cologne, DE. SSRN.

Helbing, D., Frey, B. S., Gigerenzer, G., Hafen, E., Hagner, M., Hofstetter, Y., van den Hoven, J., Zicari, R. V., and Zwitter, A. (2017). Will Democracy Survive Big Data and Artificial Intelligence? *Scientific American*. https://www.bsfrey.ch/articles/D_283_2017.pdf.

Herzing, J. M. E. and Schneider, S. L. (2014). Response Latencies and Data Quality. Cologne, DE. Presented at the General Online Research Conference (GOR 2014).

Hox, J. J. (2010). *Multilevel Analysis: Techniques and Applications*. Routledge, New York, NY, 2nd edition.

Kaplowitz, M. D., Hadlock, T. D., and Levine, R. (2004). A Comparison of Web and Mail Survey Response Rates. *Public Opinion Quarterly*, 68(1):94–101.

Knoef, M. and de Vos, K. (2009). *The Representativeness of LISS, an Online Probability Panel*. CentERdata, Tilburg, NL. https://www.researchgate.net/profile/Marike_Knoef/publication/242742051_The_representativeness_of_LISS_an_online_probability_panel/links/0f3175339ae828f081000000.pdf.

Lachin, J. M. (1981). Introduction to Sample Size Determination and Power Analysis for Clinical Trials. *Controlled Clinical Trials*, 2(2):93–113.

McCutcheon, A. L. (2002). Basic Concepts and Procedures in Single- and Multiple-Group Latent Class Analysis. In Hagenaars, J. A. and McCutcheon, A. L., editors, *Applied Latent Class Analysis*, chapter 2, pages 54–88. Cambridge University Press, Cambridge, UK.

Mohorko, A., de Leeuw, E. D., and Hox, J. J. (2013). Internet Coverage and Coverage Bias in Europe: Developments across Countries and over Time. *Journal of Official Statistics*, 29(4):609–622.

Molenberghs, G. and Verbeke, G. (2007). Likelihood Ratio, Score, and Wald Tests in a Constrained Parameter Space. *The American Statistician*, 61(1):22–27.

Pforr, K. and Dannwolf, T. (2017). What Do We Lose With Online-only Surveys? Estimating the Bias in Selected Political Variables Due to Online Mode Restriction. *Statistics, Politics and Policy*, 8(1):105–120.

Revilla, M., Cornilleau, A., Cousteaux, A.-S., Legleye, S., and de 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.

Schonlau, M., Fricker, R. D., and Elliott, M. N. (2002). *Conducting Research Surveys via E-Mail and the Web*. Rand Coperation, Santa Monica, CA.

Sterrett, D., Malato, D., Benz, J., Tompson, T., and English, N. (2017). Assessing Changes in Coverage Bias of Web Surveys in the United States. *Public Opinion Quarterly*, 81(S1):338–356.

Sudman, S. (1983). Applied Sampling. In Rossi, P. H., Wright, J. D., and Anderson, A. B., editors, *Handbook of Survey Research*, chapter 5, pages 145–194. Academic Press, San Diego, CA.

Tijdens, K. (2014). Dropout Rates and Response Times of an Occupation Search Tree in a Web Survey. *Journal of Official Statistics*, 30(1):23–43.

Tourangeau, R. (2017). Presidential Address Paradoxes of Nonresponse. *Public Opinion Quarterly*, 81(3):803–814.

Vandenplas, C. and Loosveldt, G. (2017). Modeling the Weekly Data Collection Efficiency of Face-to-Face Surveys: Six Rounds of the European Social Survey. *Journal of Survey Statistics and Methodology*, 5(2):212–232.