Data quality in probability-based online panels

Bella Struminskaya

ISBN: 978-90-393-6225-9

Print: Off Page, Amsterdam, The Netherlands — www.offpage.nl

Cover design: Carola Porzelt — www.signifi.de

© 2014, Bella Struminskaya.

All rights reserved. No part of this publication may be reproduced, stored, or transmitted, in any form or by any means, without the prior written permission of the author.

Chapter 2 was published as Struminskaya, B., Kaczmirek, L., Schaurer, I., & Bandilla, W. (2014). Assessing representativeness of a probability-based online panel in Germany. In M. Callegaro, P. J. Lavrakas, J. A. Krosnick, R. Baker, J. D. Bethlehem & A. S. Göritz (Eds.), Online panel research: A data quality perspective (pp. 61-85). West Sussex: John Wiley and Sons, Ltd. The chapter is reprinted with the permission of the publisher. No part of it may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording or otherwise, except as permitted by the UK Copyright, Designs and Patents Act 1988, without the prior permission of the publisher.

Data quality in probability-based online panels: Nonresponse, attrition, and panel conditioning

Datakwaliteit in online panels gebaseerd op een kanssteekproef:
Nonrespons, uitval en leereffecten
(met een samenvatting in het Nederlands)

Proefschrift

ter verkrijging van de graad van doctor aan de Universiteit Utrecht op gezag van de rector magnificus, prof.dr. G.J. van der Zwaan, ingevolge het besluit van het college voor promoties in het openbaar te verdedigen op donderdag 20 november 2014 des middags te 4.15 uur

door

Bella Struminskaya

geboren op 22 november 1985

te Novosibirsk, Rusland

Promotor: Prof. dr. E. D. de Leeuw
Copromotor: Dr. L. Kaczmirek

Beoordelingscommissie: Prof. dr. M. P. Couper (University of Michigan)
Prof. dr. J. W. M. Das (Tilburg University)
Prof. dr. P. A. Dykstra (Erasmus University Rotterdam)
Prof. dr. P. G. M. van der Heijden (Utrecht University)
Prof. dr. M. H. D. van Leeuwen (Utrecht University)

Acknowledgements

The guidance and support of several individuals have made the successful completion of this dissertation possible. First among them are my advisors Edith de Leeuw and Lars Kaczmirek. I am grateful for the excellent mentoring and for the support they have provided on professional and personal level. Edith and Lars are the supervisors any doctoral student could wish for. I have spent most of the time working on this dissertation in Germany, at GESIS – Leibniz Institute for the Social Sciences, with occasional trips to Amsterdam. I thank Edith for finding the time during conferences and on weekends to meet with me and discuss the dissertation, for the critical comments and encouraging remarks. If it were not for Lars Kaczmirek, I would not have started working at GESIS. I thank Lars for giving me the opportunity to pursue the doctoral degree, for his support during each step of the way, and for valuable discussions that immensely improved this dissertation.

This dissertation is a result of my work on a project GESIS Online Panel Pilot at GESIS. I thank my colleagues who have made this project possible: Lars Kaczmirek, Wolfgang Bandilla, Kai Böge, Ines Schaurer, Angelika Stiegler, Gabriele Wahlig, Kai Weyandt, as well as Siegfried Gabler and Sabine Häder. I also wish to thank my current and former colleagues at GESIS, Daniela Ackermann-Piek, Dorothée Behr, Henning Best, Tanja Dannwolf, Tobias Enderle, Matthias Ganninger, Tobias Gummer, Regina Jutz, Verena Halbherr, Timo Lenzner, Klaus Pforr, Beatrice Rammstedt, Joss Rossmann, Jette Schröder, Angela Tanner, and Jessica Walter for the stimulating intellectual and friendly environment. I especially thank Michael Bosnjak for giving me the opportunity to finish the dissertation while working on a new project and for the valuable comments and advice.

Additionally, I wish to thank Mario Callegaro and the editors of the book "Online Panel Research: A Data Quality Perspective". Their comments helped to improve Chapter 2, which has been published as part of that book. I further wish to thank Ryan DeLaney and Eike Rinke for the final proofreading and Eva Zijlmans for the Dutch translation of the summary.

While writing this dissertation, I have had a privilege to meet many individuals, who played important roles for its success. I would like to thank some of them in particular. I thank Thomas Klausch for his invaluable insight and support during the final year. I also thank Ines Schaurer. Ines, in addition to being my colleague you have become a dear friend to me in these four years it has taken me to write this dissertation. I cannot thank you enough for your comments and suggestions, and most importantly for being there through ups and downs.

My family and friends have always been supportive during the challenging, at times difficult, and rewarding endeavor of writing this dissertation. The large part of the successful completion of this dissertation I owe to my parents. They have encouraged my educational and professional aspirations knowing that it would mean my being far away from them. My dearest parents, I thank you for your love.

Bella Struminskaya

Mannheim, August 2014

Contents

Chapter 1

Introduction

1.1 Online panels as data collection tools

In recent years, online surveys have increasingly been replacing telephone and face-to-face data collection. According to the American Association of Public Opinion Research, in 2009 about 85% of studies that would otherwise have been performed using traditional methods were completed online (Baker et al., 2010). The global expenditures on online research measured as a share of expenditures in all quantitative research increased from 19% in 2006 to 35% in 2012 (Callegaro, Baker, et al., 2014). A substantial part of online research is done with online panels. The benefits of lower costs, timeliness, and access to large respondent numbers offered by online panels compared to traditional modes all contribute to the expansion of online panels. However, since online panels are a relatively new means of data collection, they go hand in hand with increased concerns about data quality.

The discussion of data quality in online panels takes into account the panel type and purposes for which the panels are conducted. A key distinction is drawn between probability-based and nonprobability panels. In probability-based panels, potential respondents are selected randomly from a frame covering the target population. Each sampled unit has a known non-zero probability of selection. Among probability-based panels, there is a further distinction between pre-recruited panels of Internet users and probability panels of the entire population (Couper, 2000). Both types start with a probability sample based on standard sampling procedures (e.g., Random Digit Dial, address-based samples, or register data) because no complete list of Internet users exists from which a sample can be drawn. The sample is then approached using a conventional mode of data collection (e.g., telephone or face-to-face interview). The difference between the pre-recruited online panels of Internet users and the full-population online panels is that in pre-recruited panels of Internet users, those who do not use the Internet are screened out during the recruitment interview and only Internet users can become panel members. In full-population online panels, non-users are given the opportunity to become panel members and are provided with an Internet access and necessary equipment (e.g., the LISS Panel and CentERpanel in the Netherlands, Knowledge Networks Panel in the USA, the German Internet Panel, and ELIPSS in France).

Nonprobability panels, on the other hand, are not based on a probability sample. There are a variety of ways to recruit members into a nonprobability online panel: placement of banners and ads on various websites, distribution of invitations via mailing lists, search engine advertisements, ads on social networking sites, snowball methods that allow panel members to recruit their friends and family members, and recruitment after an online survey or through a co-registration agreement in exchange for using online services such as news or e-commerce portals (Callegaro, Baker, et al., 2014). However, the results obtained with nonprobability panels cannot be generalized to the general population or to the Internet-using population. Nonetheless, nonprobability online panels currently dominate online panel research (Baker et al., 2010).

Multiple studies have shown that, with the exception of election studies, nonprobability online panels produce data of lower quality than do probability-based panels (see a review by Callegaro, Villar, Yeager, & Krosnick, 2014). Probability-based panels are better suited to satisfy the needs and quality demands of the scientific community. Along with the probability-based online panels that were established some years ago in the USA (e.g., Knowledge Panel of the GfK Custom Research, formerly Knowledge Networks, the Gallup Panel, the RAND American Life Panel) and the Netherlands (e.g., CentERpanel and the LISS Panel), new panels such as the German Internet Panel (GIP), and ELIPSS in France have recently emerged and even more panels are being under consideration (Das, 2012; Nicolaas, Calderwood, Lynn, & Roberts, 2014). With the increase in nonresponse over the last years in surveys using the traditional modes of data collection both in European countries and the United States (de Leeuw & de Heer, 2002), probability-based online panels seem to be an appealing alternative. However, whether the quality of the data collected by probability-based online panels is comparable to the quality of the data attained by traditional data collection methods is questionable. It has been shown that response rates in online surveys are lower than those obtained via traditional modes (Lozar Manfreda, Bosnjak, Berzelak, Haas, & Vehovar, 2008). Probability-based online panels are expensive in their recruitment and maintenance, thereby raising the question: Are the costs of constructing and maintaining probability-based online panels justified given the quality of data that can be obtained?

There are several indicators of data quality. Response rate is one of these multiple indicators of data quality and is not necessarily indicative of nonresponse bias (Groves & Peytcheva, 2008). There are several sources of bias in panel surveys: recruitment bias, nonresponse bias, attrition, and panel conditioning (Sikkel & Hoogendoorn, 2008). The multiple types of nonresponse can be the result of selected study participants' unwillingness to join the panel (recruitment bias), their failure to respond to the first interview (nonresponse bias), and their failure to respond to multiple waves or quitting the study altogether (attrition). Panel conditioning is a result of respondents' learning to answer survey questions and providing answers that no longer represent the population of study. Additionally, in some probability-based online panels an initial interview is used not only for screening and recruitment purposes but also for collecting demographic and attitudinal information on respondents, which can be further incorporated by data users into their analyses. In this case, there is also a risk of mode effects on measurement (de Leeuw, 2005, p. 239). All these potential biases influence data quality since they accumulate in the estimates that data users employ to conduct their analyses. This dissertation focuses on these five potential threats to the quality of data: nonresponse during the recruitment interview, nonresponse to the first online survey, panel attrition, panel conditioning, and the effects of the survey mode. The goal is to study the quality of data obtained from probability-based online panels. In the five studies that constitute this dissertation, the overall quality of the estimates is evaluated and the processes that influence the quality are studied: nonresponse at different stages of panel recruitment and the operation of the panel as well as measurement effects of panel conditioning and the influence of the mode. The remainder of this chapter introduces the concept of data quality that is central to the composition of this dissertation. The chapter concludes with a description of the studies in the dissertation.

1.2 The concept of data quality

Data quality is a part of a broader concept of survey quality. Central to this concept are three interrelated questions: "What is survey quality?", "How can survey quality be achieved?", and "How can survey quality be measured?" This section is organized by outlining the answers to these questions.

The concept of quality has multiple definitions. In its application to surveys, the total survey quality framework is the most inclusive (Biemer & Lyberg, 2003). The total quality framework is a survey quality paradigm that includes both the producer's and the user's perspectives. Quality is defined in terms of "fitness for use," which includes accuracy of the survey data and several other dimensions that are important for data users. Most frameworks include nine dimensions: (1) accuracy, (2) credibility, (3) comparability, (4) usability, (5) relevance, (6) accessibility, (7) timeliness, (8) completeness and (9) coherence (Biemer, 2010, p. 817). However, apart from accuracy, the number of the dimensions varies from organization to organization. Accuracy is a statistical property of an estimate and is defined as a survey estimate's deviation from its true value (ibid, p. 817). Thus, a high-quality survey provides data with the following properties: The errors are minimal; the community considers the data trustworthy; the data are comparable across time and space; they are adequately documented; they satisfy the users' needs; they are easy to access; they are released according to schedule; they contain the information needed for analyses; and they allow for combining estimates from different sources.

Survey quality (product quality) is achieved through process quality, and process quality depends on the quality of the procedures employed in an organization (Biemer & Lyberg, 2003, p. 18). For survey research, the standards are set by various professional organizations, such as the European Society for Opinion and Market Research (ESOMAR), the American Association for Public Opinion Research (AAPOR), the World Association for Public Opinion Research (WAPOR), and the International Organization for Standardization (ISO). To assure survey quality, ESOMAR has developed the "ICC/ESOMAR International Code on Market and Social Research" (ESOMAR, 2007) and the "ESOMAR/WAPOR Guide to opinion polls" (ESOMAR, 2009), among other standards. AAPOR has developed several standards and codes on professional ethics and best practices in survey research, including The Standard Definitions – Final Dispositions of Case Codes and Outcome Rates for Surveys (American Association for Public Opinion Research, 2011). Its recent advancement is the Transparency Initiative,[1] which encourages making transparency and disclosure routine for the survey profession. The regulations of the International Organization for Standardization are provided for the survey industry by the ISO-Standard 20252 "Market, opinion and social research – Vocabulary and Service Requirements."

For online panels, ESOMAR has developed a "Guideline for Online Research" (ESOMAR, 2011), "28 Questions to Help Buyers of Online Samples" (ESOMAR, 2012),

[1] http://www.aapor.org/Transparency_Initiative.htm#.U4HVExDLCYI

and the “ESOMAR/GRBN Guideline for Online Sample Quality” (ESOMAR, 2014), which was jointly developed with the Global Research Business Network (GRBN). The first two documents provide guidelines for recruitment, management, and sampling procedures of online panels, whereas the latter document focuses on sampling and the prevention of fraudulent responses. The American Association for Public Opinion Research has produced an AAPOR Report on Online Panels with a focus on data quality that summarizes current empirical results and provides recommendations for the survey practitioners (Baker et al., 2010). AAPOR has also integrated the guidelines for the calculation of the response rates for probability-based online panels and of the participation rates for nonprobability online panels (American Association for Public Opinion Research, 2011). The International Organization for Standardization has issued an International Standard ISO 26362 “Access panels in market, opinion and social research – Vocabulary and service requirements,” which supplements the ISO Standard 20252 by providing a set of definitions and service requirements for producers and users of access panels in market, opinion, and social research as well as in developing evaluation criteria for the quality assessment of recruitment, management, and sampling procedures in access panels. Many survey organizations and associations have adopted these guidelines or developed their own standards to ensure process quality in online panels (see examples in Callegaro, Baker, et al., 2014, p. 16). These standards mainly focus on nonprobability panels; they are, however, also applicable to probability-based online panels.

Another method of achieving process quality is by monitoring survey fieldwork processes and intervening to improve those processes based on the analysis of paradata, that is, data automatically gathered during the survey production process and the (computer-assisted) interview (Couper, 1998; Kreuter, 2013). In the past, paradata have been classified as meta-data (i.e., data about the survey process) but with the emergence of online surveys, which allowed for a wider collection of data not knowingly provided by respondents, the need for an independent concept became evident (Kaczmirek, 2009, p. 79). Although paradata have been widely used for research in single online surveys, paradata in online panels, which Mario Callegaro terms *longitudinal paradata*, have received little scholarly attention to date (Callegaro, 2013b, p. 272).

Quantifying or measuring the quality of survey data is achieved by measuring accuracy, which is “the cornerstone of quality” (Biemer & Lyberg, 2003, p. 24). Accuracy is one dimension of the survey quality that can be quantified, whereas other dimensions of survey quality are descriptive and therefore more difficult to measure. Accuracy is measured by estimating the Total Survey Error. The Total Survey Error is a concept that describes statistical error properties of survey estimates (Groves & Lyberg, 2010) and is outlined below in more detail.

1.3 Total Survey Error

The Total Survey Error (TSE) framework (e.g., Biemer, 2010; Groves, 1989; Groves & Lyberg, 2010) classifies errors affecting survey estimates into two major categories: errors of nonobservation and errors of observation. Errors of nonobservation result from the failure to observe the population that should be studied. They include coverage error, sampling error, and nonresponse error. Coverage error is the failure to include (or exclude) certain members of the target population in the sampling frame, which ideally is a set of all members of the target population. Sampling error is the imprecision of estimates that results from surveying a sample of respondents (which represents one possibility out of many samples that can be drawn from the sampling frame) as opposed to surveying the whole population. Nonresponse is the failure to obtain survey measures from all sampled units of the target population. Errors of observation involve measurement. Measurement error, which is the difference between the observed response and the underlying true value, arises from the response process. Processing errors occur when the responses are transferred into the database (in the case of a non-automated data collection process) or when the raw data are coded incorrectly for further analysis. Adjustment error is a failure to obtain the proper representation of the target population by correcting the observed values with statistical techniques such as post-stratification weighting or imputation.

For panel surveys, there are two additional special types of errors: panel nonresponse (attrition) and panel conditioning (Kalton, Kasprzyk, & McMillen, 1989). It has been suggested that these two additional error sources should be integrated into the TSE framework (Smith, 2011). Temporary (wave-nonresponse) or permanent nonresponse in panel surveys (panel attrition) is more problematic than is nonresponse in cross-sectional surveys. If attrition is selective, that is, the determinants of attrition are related to the statistic of interest, the resulting estimates will be biased and if it is not selective, the reduction of the sample size will result in increased variance. Panel conditioning refers to the changes in responses as a result of prior interviews. These changes may occur as a result of changing the reporting behavior or changing actual behavior or attitudes due to previous participation in the panel (Kalton et al., 1989; Waterton & Livesley, 1989). Panel conditioning can improve the response quality if respondents provide more honest or more accurate answers as a result of better understanding the survey procedure. Otherwise, when respondents learn how to manipulate the survey instrument (e.g., to recognize filter questions and skip follow-up questions) or change their behavior or attitudes after a series of interviews, a systematic bias is introduced to the data. In this case, panel conditioning can be considered a special case of measurement error. It is important to note that the goal of conventionally defined panel studies "to study changes in population characteristics and composition over time" (Kalton & Citro, 1993, p. 205) is probably shared by only a few online panels. There are various online panel designs, allowing the panels to employ both cross-sectional and repeated measures components. However, as a rule, online panels implement cross-sectional studies mostly using the longitudinal character to refresh the data from previous studies such as demographic information or the information used to select study participants for surveys on specific topics. Nevertheless, online panels are longitudinal surveys in the sense that they "collect data from the same sample elements on multiple occasions over time" (Lynn, 2009, p. 1) and are therefore prone to errors specific to panel surveys.

The errors can be introduced at each stage of the survey process, which consists of several more or less sequentially executed steps (Biemer & Lyberg, 2003; Groves, Fowler, Couper, Singer,

& Tourangeau, 2004). These steps include defining the research objectives, defining the target population and the sampling frame, determining the mode of administration, developing (and pretesting) the questionnaire, drawing the sample, collecting data from the sample members, processing and analyzing the data. Groves et al. (2004) and Groves and Lyberg (2010) connected the components of the total survey error to the inference process. Errors of nonobservation (coverage, sampling, and nonresponse) as well as the adjustment error are threats to the accurate representation of the target population. Errors of observation (measurement error) as well as misconceptions in validity (specification error: not measuring the intended concept) and processing errors threaten the accurate measurement of the survey statistic.

All errors discussed above apply to online panels. Both probability-based and nonprobability online panels share the same general building and operating stages: (1) recruitment stage, (2) profile stage, and (3) sampling for specific studies (Callegaro, Baker, et al., 2014; Callegaro & DiSogra, 2008; DiSogra & Callegaro, 2009). During the recruitment stage in probability-based panels, potential respondents are selected from a sampling frame into a probability sample and contacted by mail, telephone, or face-to-face. The contact strategies can also be combined or applied sequentially. For example, in the LISS Panel, potential respondents were sent an advance letter and then contacted by an interviewer either via telephone or face-to-face, depending on the availability of the telephone number (Scherpenzeel & Das, 2011). The recruitment may include an interview in which respondents provide information on a variety of topics and are asked about their willingness to join the online panel (e.g., the LISS Panel, GIP – the German Internet Panel[2], CentERpanel[3]).

In order to include persons without Internet access, panels may provide the necessary equipment and an Internet connection (e.g., the KnowledgePanel, the LISS Panel, GIP)[4]. Alternatively, all respondents who agree to participate in a panel may be provided with the devices necessary to participate (e.g., ELIPSS provides panel members with tablets with an Internet connection[5]). For nonprobability online panels, the recruitment stage consists of enrollment in the panel by providing some basic information (such as the respondent's name and an email address) and confirming the respondent's intent to join the panel by clicking the link sent in an invitation email — the "double opt-in" process that requires the confirmation of the intent of becoming a panel member at two distinct time points (Callegaro, Baker, et al., 2014, p. 8). During the next profiling stage, respondents are required to answer a set of survey questions, which can include demographic information, thematic profiling, or (in nonprobability panels) also questions aimed at validating the respondent's identity in order to

2 http://reforms.uni-mannheim.de/internet_panel/sample_recruitment/

3 Hoogendoorn and Daalmans (2009).

4 Another method to achieve representativeness is to allow persons who do not have access to or use the Internet, to participate in another mode. For example, in addition to the online mode, the mixed-mode GESIS Panel (www.gesis-panel.org) offers the option of mail questionnaires for persons without Internet access, low frequency Internet users, or those who refuse to participate online.

5 http://elipss.fr/elipss/recruitment/about/

prevent false registrations and the completion of the same survey more than once (Callegaro, Baker, et al., 2014). The demographic or thematic information can be used for building quota samples (in nonprobability panels) or to select subgroups of respondents for specific studies. This third stage of selection for specific studies based on certain demographic characteristics is mostly done in nonprobability online panels. However, some probability-based online panels also employ profile surveys (e.g., the LISS panel).

This dissertation focuses on probability-based online panels because only probability-based panels satisfy the scientific community's need to generalize the results to the target population although studies based on nonprobability samples may serve some study goals (cf. Baker et al., 2010). Studies on the quality of data in probability-based online panels should consider the recruitment and operation processes of such panels. In Figure 1.1, the recruitment and operating process is depicted, which is more or less common for all probability-based online panels. It shows which errors can be introduced at every stage of building and maintaining a probability-based online panel.

The first step is to define a target population. The second step is to identify an adequate sampling frame from which the sample can be drawn. If the target population is the general population, failing to include non-Internet-using respondents will result in coverage error. Moreover, since there is no available complete list of Internet users' email addresses (which would provide an ideal sampling frame), relying on other sources to build a sample can also result in coverage error (e.g., incomplete or outdated registers, failure to include cell-phone numbers when using the RDD recruitment, etc.). Sampling error occurs as a result of drawing a sample instead of investigating the total population. The third step is contacting the selected respondents for an initial (recruitment) interview. Failure to contact or refusal by successfully contacted potential respondents to take part in the interview results in nonresponse error. This is the first occurrence of nonresponse error in the recruitment process, which includes multiple decisions and can therefore lead to several stages at which nonresponse is possible. The second of these selection stages is the acquisition of the respondents' agreement to join the online panel. The third selection stage for those who have agreed to participate in the panel is to actually take part in the online survey. The loss of respondents at each of these stages prior to actual participation may make the panel biased from the start. It has been shown that this multiple nonresponse is selective and introduces bias (Hoogendoorn & Daalmans, 2009), that coverage error is more severe than nonresponse error in the interplay between coverage error and nonresponse error, (Bosnjak et al., 2013; Couper, Kapteyn, Schonlau, & Winter, 2007), and that using weighting adjustment to correct for the combination of these two error sources has its limits (Lee, 2006; Schonlau, Van Soest, Kapteyn, & Couper, 2009).

After the panel has been recruited, further error sources are panel attrition and panel conditioning. Multiple definitions for panel attrition exist (Lemay, 2009). It may encompass wave nonresponse, which can occur once or multiple times, and permanent drop-out from the panel. There are a multitude of attrition patterns in online panels (Lugtig, 2014), and persons who later drop out of the panel are different from nonrespondents at the recruitment phase (Lugtig, Das, & Scherpenzeel, 2014). On the measurement side, there is panel conditioning.

1

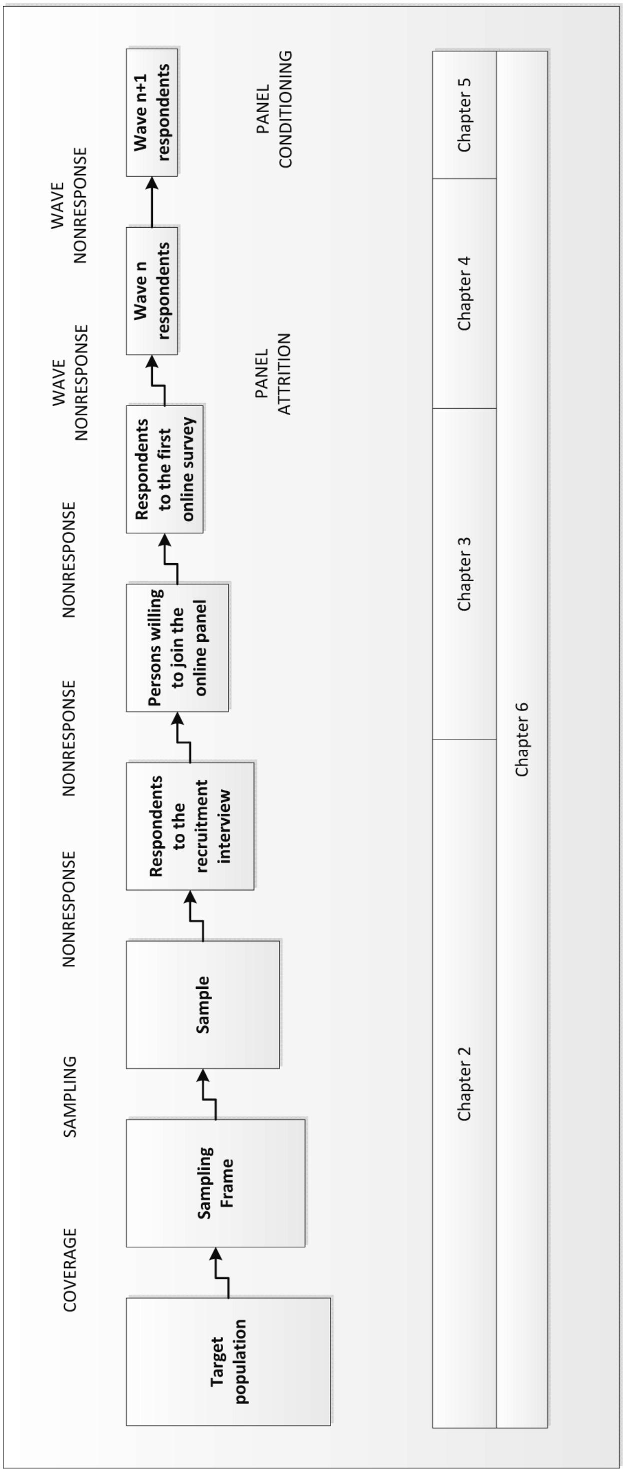


Figure 1.1. Types of survey errors in relationship to the recruitment stages of an online panel and corresponding dissertation chapters.

Panel conditioning may introduce measurement error for certain kinds of questions (Kruse et al., 2009; Toepoel, Das, & van Soest, 2009). When the recruitment has to be performed in an offline mode for probability-based panels in their current form, a mode switch between the recruitment interview and the first online panel wave is inevitable. If respondents do not feel comfortable using the Internet, the result can be nonresponse due to this mode switch. If measures from the recruitment interview are used in further analyses or if non-Internet-users are allowed to answer questions in an alternative mode, mode effects on the data quality are possible. In this dissertation, the processes of recruitment and maintenance of probability-based online panels are considered to constitute one system of data collection, which is distinguished by the features of multiple selection steps and a change of mode between the recruitment interview and further interviews. The system of data collection is an "entire data collection process designed around a specific mode," which does not have to consist of one mode only, but allows for mixing the modes (Biemer & Lyberg, 2003, p. 208).

In five studies within this dissertation, I adopt the notion of the mode system to examine the quality of the resulting estimates (Chapters 2 and 6) as well as study the processes of recruitment and panel operation, which may lead to errors in detail (Chapters 3 to 5). The overarching framework of the total survey error allows quantification through the Mean Squared Error (MSE). The MSE is calculated as the squared difference between the estimate and the parameter that it intends to estimate. The MSE can be decomposed into the systematic error (bias) and the variable error (variance) and reflects the accumulation of all errors sources (Biemer, 2010). The complete measurement of the error components has not been routinely implemented by most surveys with a few exceptions (Groves & Lyberg, 2010), and it is not performed in this dissertation. However, the studies make use of the advantage of the total survey error framework in decomposing the error sources. This dissertation aims at advancing the understanding of the causes of errors and to guide design decisions when recruiting and maintaining a probability-based panel in order to minimize the errors under given budget constraints.

1.4 Outline of the dissertation

The empirical analyses in this dissertation are based on a single data source: the GESIS Online Panel Pilot (GOPP)[6], a pilot project at GESIS – Leibniz Institute for the Social Sciences aimed at developing best practices for the recruitment and maintenance of probability-based online panels in Germany. The recruitment was based on a probability sample (RDD) of landline and mobile numbers. The last birthday method was used to identify the target respondent, who was asked to complete a short interview. Due to the pilot nature of the

[6] A description of the project and data from the selected waves can be accessed via the GESIS archive: https://dbk.gesis.org/dbksearch/GDESC2.asp?no=0058&DB=D; dois: 10.4232/1.11570, 10.4232/1.11571, 10.4232/1.11572, 10.4232/1.11560, 10.4232/1.11573, 10.4232/1.11574, 10.4232/1.11575, 10.4232/1.11576, and 10.4232/1.11577. The project was supported by the German Federal Ministry of Education and Research.

project and budgetary limitations, respondents who did not use the Internet could not be provided with the necessary equipment and Internet access and were screened out at this stage. At the end of the interview, all Internet users were asked to provide their email address in order to participate in the online panel. In the case of agreement, the respondent was sent an email invitation with a link to the survey once a month for eight months. The online surveys included a broad range of topics, such as work, family life, religion, social networks, and environment. Various methodological experiments were implemented in the panel, including an incentive experiment and experiments on panel conditioning.

This dissertation is a collection of studies that examine the quality of data collected with this offline-recruited probability-based online panel. Each chapter studies a different aspect of data quality. The chapters are written as individual papers that can be read independently from each other, which results in an overlap in the sections which describe the data used for analysis.

Figure 1.1 shows how different chapters in this dissertation correspond to the recruitment and operating steps of a probability-based online panel and the error sources associated with each step. At each step of the recruitment process, the estimates accumulate the errors introduced at the previous stages, thereby affecting the resulting data quality.

Chapter 2 provides a detailed description of the recruitment process and compares the estimates from the online panel and the recruitment interview with two external benchmarks: the German General Social Survey ("ALLBUS") and the German sample of the European Social Survey (ESS). The point estimates for demographic and attitudinal variables are compared across the surveys. Furthermore, the extent to which the data from the online panel produces comparable results when it is used for regression models is analyzed. The potential of weighting to correct for selection biases is also discussed. Chapter 2 provides the most detailed description of the GESIS Online Panel Pilot.

Chapter 3 and the next two chapters focus in more detail on the processes that lead to potential biases. In Chapter 3, I study two components of nonresponse to the online panel: refusal of respondents during the telephone recruitment interview to join the online panel (nonwillingness) and nonresponse to the first online survey. It is examined how various respondent characteristics and survey design features, including incentives and the fieldwork agency, influence the likelihood of response at each of these two stages. In line with viewing offline-recruited probability-based online panels as systems of data collection, respondent characteristics include predictors, which account for respondents' affinity toward the Internet, experience with the Internet, and experience with online surveys. It is studied how this mode-specific nonresponse varies between the stages of willingness and actual participation.

Chapter 4 studies attrition, which is defined as nonresponse occurring after respondents have joined the panel and participated in the first online questionnaire. Applying the benefit-cost-theory of survey participation, which emphasizes the intrinsic motivations of survey participation, respondents' survey experience (intrinsic) is contrasted with offered incentives (extrinsic) motivational factors. The main question of this chapter is whether incentives inhibit panel attrition and if so, whether they serve as compensation for a negative experience or whether they serve as an additional benefit. It further differentiates between several attrition groups based on their response patterns. The distinctive feature of the analysis in this chapter is the focus on time-varying covariates of survey experience.

Chapter 5 focuses on panel conditioning. It is argued in this chapter that viewing panel conditioning exclusively as a negative phenomenon that threatens data quality is not comprehensive. Learning the survey process, which can lead to both advantageous and disadvantageous effects on the data quality, is studied. Two field experiments are aimed at uncovering panel conditioning in the online panel. The problem of confounding factors is approached with propensity score weighting. Furthermore, I analyze paradata to rule out alternative explanations of panel conditioning.

Chapter 6 draws on the notion of the system of data collection and compares data that were collected during various waves of the online panel with two cross-sectional reference surveys: the German General Social Survey ("ALLBUS") from the years 2010 and 2012. This chapter seeks to single out the mode system effect (the closest to the pure mode effect in the practical circumstances of data collection) by comparing three systems of data collection. Two of these systems are extremely similar (the recruitment and interviewing mode was face-to-face, fieldwork was carried out by the same fieldwork agency), allowing to draw a number of conclusions about the quality of online data. For the analysis, a set of questions with identical wording is chosen, and the differences in sample compositions are controlled for by weighting. It differentiates between factual and attitudinal questions. For attitudinal questions, more pronounced effects are expected than for factual questions.

Chapter 7 concludes the dissertation by summarizing the major results and discussing the directions for future research. The results of the five studies in this dissertation will be useful for researchers who plan to build probability-based online panels of Internet users or probability-based online panels of the general population. Furthermore, the results of specific studies, for example, presented in Chapter 2 and in Chapter 6, might be useful for existing interviewer-administered panel surveys that consider changing the mode of administration to the online mode.

Chapter 2

Assessing representativeness of a probability-based online panel in Germany

This chapter was co-written with Lars Kaczmirek, Ines Schaurer
and Wolfgang Bandilla and was published as: Struminskaya, B.,
Kaczmirek, L., Schaurer, I., & Bandilla, W. (2014).
Assessing representativeness of a probability-based
online panel in Germany.
In M. Callegaro, P. J. Lavrakas, J. A. Krosnick, R. Baker,
J. D. Bethlehem & A. S. Göritz (Eds.),
Online panel research:
A data quality perspective (pp. 61-85). West Sussex:
John Wiley & Sons. Reprinted with a permission from the publisher.

2

Abstract

The growth of online survey research has led to an increased demand for probability-based online panels. Several of such panels are established in Europe and in the USA. While probability-based online panels are being used by scientific institutions to collect data and make inferences about the target population, questions about the quality of such data continue to be raised. In this chapter, we assess the quality of an offline-recruited probability-based online panel of Internet users in Germany. First we report the key performance measures of the telephone recruitment and online surveys. In the second step of the quality assessment, we compare our data to other surveys. As benchmarks we use two population surveys: the German General Social Survey (ALLBUS 2010) and the German sample of the fifth round of the European Social Survey (ESS 2010). Both of these sources contain the information on private Internet usage and thus allow us to compare the estimates from our panel with the estimates calculated for subsamples of Internet users from the reference surveys. Both demographic and attitudinal measures are considered. We assess the feasibility of post-stratification weighting to correct for noncoverage and nonresponse. Additionally, we assess the comparability of the three surveys in modeling social phenomena. This chapter provides insight into the quality of data collected via online panels and discusses the efficiency of probability-based online panels as means of data collection for scientific purposes.

2.1 Probability-based online panels

Advances in online survey research methods have led to substantial increase in the use of online panels, mainly of non-probabilistic nature. According to the AAPOR Report on Online Panels, in 2009, about 85% of research, which might otherwise have been performed by traditional methods, was done online (Baker et al., 2010).

In academia, cost-efficiency benefits of web-based interviewing are paired with concerns about the quality of data gained from non-probability online panels. Increasing nonresponse in traditional methods and research showing that claims of representativeness by non-probability online surveys are hardly confirmed, led to an increased demand for probability-based online panels. Among the established online panels based on probability samples worldwide are the LISS Panel and the CentERpanel (in the Netherlands), RAND American Life Panel, Gallup Panel, and KnowledgePanel of GfK Custom Research, formerly Knowledge Networks (in the United States). In Europe, new online probability-based panels are emerging, examples include ELIPSS in France and GIP in Germany (Das, 2012).

Studies comparing probability and non-probability online panels continue to show that probability-based online surveys yield more accurate results (Chang & Krosnick, 2009; Scherpenzeel & Bethlehem, 2011; Yeager et al., 2011). Probability-based panels and their data are nonetheless constantly being evaluated because, like all surveys, they are prone to the threats to representativeness.

Of particular importance are errors of non-observation: first, noncoverage or failure to include all eligible members of the target population in the sampling frame and, second, unit nonresponse or failure to gather data from all members of the sample. Undercoverage becomes less of a concern with the inclusion of cell-phone only households in the Random Digit Dial (RDD) telephone sampling frames for telephone recruitment and increased use of address-based sampling for face-to-face and mail recruitment strategies.

Nonresponse at a unit level in online panels poses a problem since online panels usually employ multi-step recruitment protocols in which the initial contact and panel membership request is made face-to-face or via telephone (Vehovar, Lozar Manfreda, Zaletel, & Batagelj 2002). The second threat to accurate representation of the target population is wave-nonresponse or attrition when persons who took part in an initial online interview fail to do so for one, several or all the following waves. If factors influencing wave nonresponse or attrition are correlated with the outcomes of interest, losing these respondents can lead to biased data.

One way to assess representativeness of the data from online panels or online surveys in general is to compare estimates with external benchmarks. Usually government statistics or data from surveys with high response rates are used as a reference. Both sources have their advantages and flaws. Government statistics do not commonly go beyond demographic information, whereas researchers and clients are more interested in opinions and attitudes. Survey data from face-to-face or telephone surveys contain non-demographic information but are themselves subject to nonresponse.

In this chapter we employ the latter strategy. We compare data from the probability-based online panel of Internet users in Germany to data from two high-quality face-to-face surveys. After the description and assessment of the recruitment process, we study sample composition employing several demographic measures as well as including attitudinal variables in our analysis. Finally, we address the challenges of having two benchmark surveys, which have variation within them, and consequences for drawing inferences from our data.

2.2 Description of the GESIS Online Panel Pilot

2.2.1 Goals and general information

The GESIS Online Panel Pilot (GOPP) was a methodological project primarily aimed at developing effective recruitment and panel-maintaining strategies of a probability-based online panel in Germany for the academic research purposes of social and political scientists.

Respondents were recruited into the panel via telephone using a probability sample. To avoid the issues of decreasing landline coverage, cell phone numbers were included in the sample. With the project having mainly methodological purposes, persons who did not use the Internet were screened out during the recruitment interview. Internet users were asked at the end of the recruitment interview to provide their email address to participate in an online panel. Those who agreed were sent email invitations to online surveys of durations between 10–15 minutes every month for eight months in total.

GESIS Online Panel Pilot is a longitudinal panel in the sense that it collects data from the same sample units on multiple occasions over time (Lynn, 2009). Every monthly questionnaire had a leading topic and included some demographical questions in order not to burden the respondent with all factual questions in a single survey. A small proportion of repeated measures were collected over the course of the panel. Those measures were mainly included for methodological reasons rather than to study social change.

2.2.2 Telephone recruitment

The target population for the GESIS Online Panel Pilot were German-speaking persons living in the Federal Republic of Germany aged 18 and older, who used the Internet for non-work-related purposes. With the growing mobile-only population in Germany, it was decided to use the dual frame approach for the telephone recruitment. Samples were drawn independently from landline and cell phone frames. The goal was to have a final sample with 50% of eligible cell phone numbers and 50% of eligible landline numbers. In order to handle the overlap between two frames, we calculated the inclusion probabilities for target persons using formulas of Siegfried Gabler, Sabine Häder and their colleagues under the assumption that two samples are independent (Gabler, Häder, Lehnhoff, & Mardian, 2012). The inclusion probabilities account for the sizes of the samples and frames in both cell phone and landline components as well as the number of landline and mobile numbers at which a respondent can be reached, and for the landline part the size of the household. Within the landline part of the dual frame, the contact person was asked to state the number of adults residing in the household and the target person was the person with the most recent birthday. In the cell phone part of the frame no selection procedures were implemented, since cell phone sharing in Germany is estimated to be about 2% (Gabler et al., 2012). For both landline and cell phone samples persons contacted on institutional telephone numbers or numbers used for non-private use were excluded from the sample. In addition, respondents were asked to provide the total number of landline and cell phone numbers at which they could be reached. The questionnaire employed a complex routing scheme to take into account various combinations of landline, cell phone, and special telephone contracts available in Germany (e.g., landline numbers with cell phone contracts). The project was executed in three sequential parts, each of those employed the strategies, which proved to be efficient during the previous part or corrected for deficiencies of the previous part. The first part, preceded by a short pretest, started the recruitment in February 2011 and continued until April 2011. Recruitment of the second part was done between June and July 2011. The recruitment for the third part lasted from June 2011 to August 2011 (for differences between study parts, see Table 2.1). Up to 15 call attempts were made to reach the number. In the second and third part of the recruitment the number of call attempts was reduced to 10.

After a target respondent was identified, he or she was asked to participate in a short interview (about 10 minutes). At the beginning of the interview respondents were asked about non-work-related Internet use. For non-users the recruitment interview ended at this point. It is good practice to collect information about non-users as well. However, as studying the differences between Internet users and non-users was not a central part of the project, such information was only collected for a subsample. Internet users were asked

Table 2.1. Overview of the features of the GESIS Online Panel Pilot.

Feature	Study 1	Study 2	Study 3
Fieldwork start	February 2011	June 2011	July 2011
Fieldwork agency	In-house, own telephone lab	In-house, own telephone lab	External, specialized on market research surveys
Internet-screener	The very first question with target sample unit	Preceded by a general attitudinal question	Preceded by a general attitudinal question
Incentive experiment	- 10 EUR plus 20 EUR bonus for completion of all 8 questionnaires, - 5 EUR plus 20 EUR	- 5 EUR, - 2 EUR, - 0 EUR	- 5 EUR, - 2 EUR, - 0 EUR
Announced panel duration	8 months	No limitation	No limitation
Length of the recruitment interview	Regular version	- Regular version (90%), - Shorter version (10%)	- Regular version (90%), - Shorter version (10%)
Experiment on household income question	Household income question asked in 50% of cases	Household income question asked in 50% of cases	Household income question asked in 50% of cases

several attitudinal questions about living in Germany, leisure activities, basic demographic information, frequency of Internet use and online survey experience, as well as questions on household composition and number of landline and mobile numbers.

At the end of the interview, respondents were requested to provide an email address to participate in the online panel. Interviewers were instructed to describe the concept of the online panel and state the amount of incentive the respondent would receive as a result of participating in each online survey. Incentive conditions varied experimentally from 0–10 Euros per completed online questionnaire.

Apart from providing the incentive information, interviewers also answered any additional questions which respondents asked about the panel. This part of the interview was less standardized, interviewers were encouraged to perform it as a dialogue with the respondent and to apply tailoring strategies to maximize the recruitment rate. With interviewers making effort to persuade respondents to participate in the online panel, hesitant respondents were offered the opportunity of a lagged decision. This meant that individuals having concerns about the scientific nature of the study were referred to the project homepage and followed up by call-backs a few days later.

To explore opportunities of increasing the recruitment rate, two experiments were integrated into the interview process. The first experiment tested the effect of asking a question about household income on the recruitment success. In a split-half experiment we asked half of the respondents to report their monthly household income, providing bracketed categories for their answers. The second half of respondents were not asked about their income.

The second experiment was implemented in the second and in the third recruitment parts. The experiment varied the announced length and content of the telephone interview. To test the effect of the interview length on recruitment success, for the shorter version, the recruitment interview was reduced to an introductory question about living in Germany, basic demographic information, and the panel participation request.

For certain conditions such as when a target person refused to participate in telephone surveys in general, multiple appointments failed before, and/or a target person seemed to be near an interview break-off, the option of only providing an email address was given to the respondent. Interviewers asked about Internet use, briefly described the project and probed for an email address. No demographic information was collected during the interview. For those respondents who joined the online panel, the required additional demographic information and information needed for calculation of design weights was collected in the second online survey.

2.2.3 Online interviewing

Respondents who agreed to participate in the online panel and provided their email address were sent an email invitation which contained a link to the first online survey. The invitation also included incentive information for experimental groups in non-zero incentive conditions.

After clicking on the link in the invitation email, respondents were redirected to the online survey. Each online questionnaire started with a welcome screen, the content of which stayed unchanged during the entire course of the panel except for announcing the number of the survey respondents were about to complete. The welcome screen contained general information about GESIS as the sponsor of the survey, a link to the data protection policy, and contact information.

In the first online survey respondents were fully informed about data protection issues. In all following surveys a shorter version of the data protection clauses was employed with a link to the full text. The first screen in all online surveys was followed by a page with instructions on filling in the questionnaire.

Each of the monthly online questionnaires had a main topic which most of the questions addressed. Additionally, demographic questions were spread over questionnaires 2–4. We used this approach instead of starting with a profile survey to make the experience of filling out the first online survey as pleasant and interesting as possible. The lower respondent burden in the beginning aimed to increase the likelihood of respondents staying in the panel.

The main topics of online questionnaires were:
- Questionnaire 1: Multi-topic introductory wave
- Questionnaire 2: Education and employment
- Questionnaire 3: Family life
- Questionnaire 4: Religion and values
- Questionnaire 5: Environment, ecology
- Questionnaire 6: Social networks
- Questionnaire 7: Politics
- Questionnaire 8: Multi-topic (with a focus on gender roles and personality).

Most of the questions used in the online questionnaires were originally conducted as part of other German and international surveys. The panel had a strong focus on replicating questions for two reasons: first, rather complex constructs were implemented to demonstrate their feasibility in an online setting; second, to assess data quality, comparisons with external benchmarks were planned.

For these reasons, a substantial part of questions originated from the German General Social Survey ("ALLBUS") 2010 and the German version of the European Social Survey 2010. Other questions were replicated from a wave of the German Socio-Economic Panel (GSOEP 2008), the Eurobarometer, and the International Social Survey Programme (ISSP). Generally, these questions were implemented using the original question wording. In some cases the wording was adjusted to match the self-administered mode.

Apart from questions on specific topics, each questionnaire ended with the same set of questions to evaluate the current survey. Respondents were also asked whether they had interrupted the completion of the survey and if so, how long the distraction lasted. There was a text field for comments offered each month at the end of the questionnaires. Providing comments was not obligatory for the respondent. Several paradata, that is, data about the process of survey production (Kreuter, 2013), were collected in the course of the online survey. These included, for example, the duration of questionnaire completion, response latencies, and browser information.

The last page of every questionnaire contained a thank you note and information about the amount of incentive respondents received for completing the survey. The control group, which received no incentives, was presented with a thank you note only. Respondents could redeem incentives in various ways: paid to their bank account in Euros or as an Amazon-voucher, in which case no bank account information was required. In the study parts 2 and 3, participants could also choose to have the money donated to various charity organizations.

The panel, as well as telephone recruitment interview, included experiments about survey methodology. One of those experiments was aimed to study panel conditioning, the phenomenon when the act of participating in surveys changes attitudes and behavior or the way respondents report attitudes and behavior (Kalton, Kasprzyk, & McMillen, 1989). It employed rotating questionnaires: while some respondents received a specific questionnaire in the beginning of their panel participation, the same questionnaire was completed by others after having completed a few surveys.

Another experiment aimed to capture changes in social desirable responding. It included a repeated measurement design for a subsample of the online respondents recruited during the third recruitment round.

In case of no answer by the respondent for a week, a reminder was sent to the respondent with a text and a link to the online survey. Overall, between the waves of the online surveys a respondent could receive a maximum of three reminders, one each week, if he or she failed to start the online survey. After a month, the respondent received an invitation to the next online survey; however, a chance to complete non-answered questionnaires was given in the course of the entire eight months. For this purpose, respondents could log in to the panel with the authentication information they received after completion of the first online survey.

Panel members who decided to stop participation could opt out by sending a request to be removed from the panel. In this case they received no further invitations or reminders.

2.3 Assessing recruitment of the Online Panel Pilot

In this section we describe the response metrics for the telephone recruitment and online participation. The presentation of the response rates is twofold. We first present the metrics from the telephone recruitment and then proceed to the online part of the panel. Although all metrics were also calculated separately for each of the three parts, we only report the average values over three main study parts. These values are also the basis for our analysis later in the chapter. The following response metrics refer to the formulas by Callegaro and DiSogra (2008) and to the AAPOR final disposition codes for RDD telephone surveys (American Association for Public Opinion Research, 2011).

For completed telephone interviews, irrespective of whether a respondent joined the panel or not, the overall response rate AAPOR RR3 was 17.8%, with the landline sample resulting in 18.5% and the mobile sample reaching 17.1%. The response rate is comparably high to other telephone surveys in Germany and also to surveys in the US. The Pew Research Center estimated an average AAPOR RR3 of 9% for a typical telephone survey (Kohut, Keeter, Doherty, Dimock, & Christian, 2012). Our sample consisted of more than 95000 telephone numbers. The contacts were nearly evenly distributed among landline and cell phone respondents.

Respondents to the telephone interview were defined as eligible if they used the Internet for non-work-related purposes and were, among other criteria, of age (18+) and able to participate in a German-language telephone interview. Respondents were eligible if they used the Internet at home or outside of the home, but were ineligible if they only used the Internet at work. Overall, 4840 telephone interviews were completed, of which 3514 respondents confirmed Internet-usage (72.6%). Of those, 1665 respondents were willing to participate in the panel and provided valid email addresses which could successfully be used to send out the link to the first online survey. From those, 1010 persons started the first survey and 934 persons completed it.

From the available figures, we can calculate the response metrics. The recruitment rate is the proportion of respondents who provided initial consent over all eligible respondents (i.e., among other criteria, using the Internet for private purposes). Here, we follow a conservative approach and only count those respondents as having provided their consent who not only said that they would like to join the panel but also provided a valid email address.

Recruitment rate (RECR) = 9%.[7]

As part of quality assurance, email addresses were screened while it was still possible to re-contact respondents via telephone so that they could correct any misspellings. If the welcome email which was sent to respondents shortly after the interview did not go through, interviewers called the respondents again to correct the email address. Roughly 10% of all emails were misspelled, either accidently or on purpose in the first attempt and thus proved non-working. All respondents with a misspelled email address were called back in order to correct the email address. Possible outcomes of these calls were correction of the email address, refusal to provide the interviewer with an email address thus refusing to participate in the panel or non-contact. If corrected emails still did not work, CATI fieldwork managers could decide if further follow-up was needed for further clarification. Those cases for which still no clarification was possible were treated as refusals in response rates calculations. The respondents, who provided a working email address, could still reconsider their choice and not complete the first online questionnaire. Therefore, a respondent was only defined as an active panel member after having completed the first online questionnaire. The corresponding metric was termed profile rate (Callegaro & DiSogra, 2008), which is the proportion of all participants who completed or partially completed the first online questionnaire over all invited persons. Since the panel did not use a profile survey as a first survey but distributed the collection of demographic information across the first few online questionnaires, the profile rate is identical to the completion rate of the first survey.

Completion rate (COMR, started the first online survey) = 60.7%.
Completion rate (COMR, completed the first online survey) = 56.1%.

[7] Please refer to Appendix 2 for detailed information about the calculation of the recruitment rate.

The cumulative response rate follows directly from multiplying the recruitment rate (RECR) and the completion rate of the first survey.

Cumulative response rate 1 (CUMRR1) = 9% x 56.1% = 5%.

When considering this last figure, it is necessary to keep in mind that the panel study was comprised of three separate studies, all of which were geared towards assessing the effect of different design features. Naturally, this included many suboptimal recruitment conditions (e.g., a control group without any incentive), which lowered the overall response metrics. In the next section we assess the data quality of the panel by comparing estimates derived from respondents' answers during the recruitment process and the first online questionnaire to estimates from other surveys.

2.4 Assessing data quality: Comparison with external data

In the remainder of this chapter we assess the quality of data, which was collected in the GESIS Online Panel Pilot. Data quality is an aspect of representativeness, which is determined by the quality of recruitment procedures as well as online nonresponse and online response quality. In the previous section we described the recruitment success, reporting the recruitment rates and online response rates. Response rates are not necessarily indicative of nonresponse bias (Groves & Peytcheva, 2008) and nonresponse analyses are not the focus of this chapter. Instead we define data quality as the goodness of our estimates. Estimates are the end product of a survey, which researchers perform their analyses with, and whose results are used for decision- or policy-making. Possible biases resulting from the telephone recruitment, unwillingness to join the panel and online nonresponse are all comprised in the online estimates.

Assessing representativeness by comparing estimates from an online survey to official records or other external sources of information where one particular statistic is present is not uncommon (Bandilla, Kaczmirek, Blohm, & Neubarth, 2009; Chang & Krosnick, 2009; Scherpenzeel & Bethlehem, 2011; Smith, 2003; Yeager et al., 2011). The gold standard is to compare online survey results to administrative records and/or population and government registers. These sources are not free of problems: they are difficult and costly to obtain in practice, can be outdated, relate to different reference population, and contain different operational definitions (Lynn, 2008). An alternative course of action is to use estimates from offline general population surveys with high response rates.

In our case, population registers do not contain the information about the Internet usage: the proportion and the characteristics of the Internet-using population in Germany are themselves subject to estimation (Destatis [the German Federal Statistical Office], 2011). This does not allow us to use official sources since our target population is restricted to the Internet-using population. Moreover, attitudinal measures are usually not included in official statistics reports, what makes them less valuable for assessing the quality of the online panel if one wishes to look beyond demographic measures.

We use two reference surveys to assess data quality: the German General Social Survey (ALLBUS) and the German part of the European Social Survey (ESS). They both include information on private Internet usage and had fieldwork performed within a similar timeframe as the fieldwork of the online panel. The reasons for treating these surveys as benchmarks in terms of data quality are (1) relatively high response rates and, more importantly, (2) selective nonresponse in the multistep recruitment of the online panel.

2.4.1 Description of the benchmark surveys

2.4.1.1 German General Social Survey (ALLBUS)

The German General Social Survey is a cross-sectional, interviewer-administered, face-to-face survey on attitudes, behavior, and social structure in Germany. It was first conducted in 1980 and has been conducted by GESIS every two years. For our analysis, we use the 2010 data. In 2010, ALLBUS employed a two-stage disproportionate random sample of all persons who resided in private households in the Federal Republic of Germany on the day of the interview and were born before 1 January 1992 with oversampling households in East Germany. In the first sample stage, municipalities in western Germany and municipalities in eastern Germany were selected with a probability proportional to their number of adult residents; in the second sample stage, individual persons were selected at random from the municipal registers of residents. The reported response rate in 2010 was 34.4 percent[8], which was not unusual for Germany following the European trend of decreasing response rates (de Leeuw & de Heer, 2002). The initial number of interviews in ALLBUS 2010 was 2827. We excluded from our main analysis all persons, who did not use the Internet, which decreased the sample size to 1869.

2.4.1.2 German subsample of the European Social Survey (ESS)

Like our first reference survey, the European Social Survey is a cross-sectional face-to-face biennial survey. It was first conducted in 2002. ESS is a multi-country study, which includes over 30 countries. It is funded by the European Commission, European Science Foundation, and national funding councils in participating countries.[9] For our analyses we use the German subsample of the 5th ESS round conducted in 2010. The sampling procedure for Germany is the same as it is for the ALLBUS, however, the target population includes individuals residing in private households aged 15 and older. For our analysis we further excluded all individuals aged 15–17 and the non-Internet-using population. The resulting sample size of the ESS was 2038.

Further information about both reference surveys and the GESIS Online Panel Pilot is presented in Table 2.2.

---

8 http://www.gesis.org/en/allbus/study-profiles/2010/

9 http://www.europeansocialsurvey.org/ and http://ess.nsd.uib.no/ess/round5/

2

Table 2.2. Sample description for GESIS Online Panel Pilot and two reference surveys.

Survey feature	GOPP	ALLBUS	ESS
Sample source	Adjusted Random Digit Dial sample (landline and cell phones)	Two-stage probability: municipalities with the probability proportional to the population size of municipality, individuals from municipal registers (random)	Two-stage probability: municipalities with the probability proportional to the population size of municipality, individuals from municipal registers (random)
Number of sampling points (East and West Germany)	N/A	162	168
Target population	Individuals residing in households in Federal Republic of Germany aged 18 and older who use the Internet	Individuals residing in households in Federal Republic of Germany aged 18 and older	Individuals residing in households in Federal Republic of Germany aged 15 and older
Prenotifier	N/A	Advance letter, advance brochure	Advance letter, advance brochure
Recruitment mode	CATI	CAPI	CAPI
Mode of data collection	Online, CATI	CAPI	CAPI
Response rate	CATI AAPOR3: 17.8% Online Completion rate first online survey: 55.6%	34.4%	30.5%
Field dates	February 2011-May 2012	May-November 2010	September 2010-February 2011
Incentives	Experimental groups of 0, 2, 5 EUR, 5+bonus 20 EUR, 10+bonus 20 EUR	0, 10 EUR (Interviewers' tailoring)	20 EUR (cash), conditional upon completion
N	794[a]	2827	3031
N Internet users (18+)	794[a]	1869	2038
Percentage Internet users, weighted	71.5%[b]	67.1%	72.5%

[a] Only regular (nonexperimental) telephone interviews, overall 934 respondents completed the first online survey.
[b] For the Online Panel only unweighted proportion of Internet users is available, since non-users were not interviewed and thus answered no questions on the basis of which weights were constructed.

2.4.2 Measures and method of analyses

We examined demographic and attitudinal measures. Among demographic variables were gender, age, educational attainment, legal marital status, employment status (whether working for pay or not), immigration background. Operationalization of the employment status is not exactly the same in ESS and ALLBUS with ESS asking about being in paid work during the last 7 days and ALLBUS differentiating between full-time, part-time, less than part-time and not working. Not stating a reference period, the ALLBUS operationalization may slightly overestimate those in paid employment compared to the ESS. The operationalization in GOPP is comparable to the ALLBUS. The binary variable "working for pay" contrasts those in paid work regardless of the type of employment.

Attitudinal variables used for comparison were political interest, satisfaction with the government, generalized trust, self-rated health status[10], assessment of the state of health services in Germany, rating of current state of German economy, rating of German economy in one year, rating of own respondent's financial situation, rating of own respondent's financial situation in one year, and general life satisfaction.

In order to eliminate the effect of attrition in the panel data, we only consider measures collected during the recruitment interview and during the first online questionnaire. Data collection for the first online survey was mainly finished in August 2011. Information on the specific mode in which the variables were collected can be found in Table A.1 in Appendix A.

Data quality is assessed by comparing the estimates from GOPP to the two benchmark surveys (Table 2.3). Depending on the scaling of the variable, we compared proportions or means. Statistical significance of difference between pairs of estimates from each benchmark survey was tested using *t*-tests with the hypothesis that a difference between two proportions/two means equals zero.

Data in both benchmark surveys were weighted using the design weights which accounted for unequal probabilities of selection. For GOPP, the weights were constructed by GESIS. The reference surveys provided the respective weights. The ALLBUS documentation states that since it is based on a sample of individuals drawn from the municipal registers of residence, it needs no weighting when conducting individual-level analysis. However, individuals residing in former East Germany are oversampled. The ALLBUS design weights correct for that.[11] Design weights for the European Social Survey are computed for each country depending on the number of stages in the sampling design. For Germany, this number equals 2. The weights are calculated as the inverse of the combined probability of being selected during each stage.[12] For both surveys the weights are rescaled so that the sum of final weights equals the sample size. The GOPP weights correct for the unequal inclusion probabilities as described in section 2.2.2. The design weight was calculated as the inverse of the inclusion probability for each respondent. For the inclusion probability we used the following formula: $\pi_i \approx k_i^L \frac{m^L}{M^L} \times \frac{1}{z_i i} + k_i^C \frac{m^C}{M^C}$, where k is the number of telephone numbers where a respondent can be reached, m is the sample size of numbers, M equals frame size,

10 Scale labels for the variable self-assessed health differ somewhat between ALLBUS and ESS. Labels used in ALLBUS were: very good, good, satisfactory, poor, bad; labels used in ESS were: very good, good, fair, bad, very bad. For comparison a dummy variable was constructed which contrasts reports of very good and good health taking the value of 1 and other values when it takes the value of 0. GOPP used the same operationalization as ALLBUS.

11 Sample design and weighting used in ALLBUS: http://www.gesis.org/fileadmin/upload/dienstleistung/daten/umfragedaten/allbus/dokumente/Weighting

12 Sampling design and weighting procedures are described in the ESS Data Documentation Report for Round 5: http://ess.nsd.uib.no/ess/round5/

and z is the size of the household the target person resides in. L and C refer to the landline and cell-phone components (Gabler et al., 2012).

As the Online Panel estimates are obtained after several stages of nonresponse, which can affect the resulting differences between the three surveys, we use post-stratification weighting to correct for this. Ideally, we would have taken Internet population benchmarks to calculate post-stratification weights. However, neither German Census which was carried out in 2011, nor the Microcensus (the official representative 1% sample survey of the German population) contains indicators of Internet use. The study by the German Federal Statistical Office "Information and communication technologies (ICT) in households,"[13] which provides Internet usage benchmarks for Germany, uses a quota sample, which makes it unsuitable for calculating weights as the quality of the sample cannot be assessed. None of the other studies with a specific focus on Internet which we reviewed (a listing of which can be found in Kaczmirek and Raabe, 2010) meets the criteria for one or the other reason (e.g., usage of telephone surveys to collect information about Internet users but exclusion of cell phone users). For these reasons, we use ALLBUS to calculate post-stratification weights. Specifically, we used the distributions of age, gender, and education of the subgroup of Internet users to construct the weights. Ideally we would report the results for ALLBUS and ESS using post-stratification weights as well. However, ALLBUS does not provide post-stratification weights and ESS post-stratification weights were not available at the time of writing. Therefore, unweighted data from ALLBUS was used to construct the weights. As ALLBUS is itself a sample survey prone to nonresponse, one needs to be careful interpreting the results. We address this issue in the results section.

On the level of individual estimates, if respondents refused to answer the question or provided a "don't know" response, those cases were not included in the calculation of means or proportions. For design weights in GOPP we imputed missing values for variables used to calculate weights (number of landline and cell phone numbers where a respondent can be reached, number of household members over 18 years of age, and other variables) with means of respective variables, which were calculated based on the entire sample. For the two reference surveys weighting variables did not contain missing data.

2.5 Results

2.5.1 Demographic variables

A straightforward procedure of comparing estimates from the panel to the reference surveys and applying t-tests for statistical significance of the pairwise comparisons for demographic variables showed that the panel did not differ from the two benchmark surveys on gender composition. It further did not differ substantially with respect to age groups. Although respondents in the panel were somewhat younger than in the reference surveys (mean age

13 https://www.destatis.de/EN/Meta/abisz/IKTPrivateHaushalte_e.html

in GOPP=41.1, in $ALLBUS_{(A)}$=42.6, $ESS_{(E)}$=43.0; t_A=-2.20, p_A=0.028, t_E=-2.76, p_E=0.006) and overall age distribution of ESS differed from GOPP (χ^2_A= 8.41, p_A=0.078; χ^2_E=12.95, p_E=0.012)[14], when looking at the groups, the only significantly different age group in GOPP was the one of 50–64 years. It differed from ESS by 4.5 percentage points.

Educational level was the one variable among demographics, which demonstrated the largest differences with the reference surveys. Particularly overrepresented were higher-educated Internet users: the difference with ALLBUS was 16.2 percentage points, the difference with ESS was 15.2 percentage points. Consequently, the group comprising low educational level was highly underrepresented with a 10.5 percentage point difference from ALLBUS and a 7.3 percentage point difference to ESS (χ^2_A= 88.13, p_A<0.001; χ^2_E=71.70, p_E<0.001). The proportion of respondents with low educational level also differed between the two reference surveys: the estimate from ESS differed from ALLBUS by 3.2 percentage points (p=0.024). That is the first of several differences between the reference surveys, which will be discussed later in the section.

Three other demographic variables demonstrated dissimilar patterns across reference surveys but all somehow differed in GOPP from either or both of the comparing surveys. The proportion of married respondents was significantly higher in both reference surveys compared to GOPP (p<0.001). Fewer individuals with immigration background were found in the panel compared to ALLBUS and compared to ESS. Less respondents working for pay were found in the panel than in ALLBUS but no such differences were present when compared to ESS. This is particularly interesting since question wordings varied in ALLBUS and ESS and the ALLBUS wording was used in the panel. Therefore it was reasonable to expect differences both between ALLBUS and ESS (which is the case) and between GOPP and ESS.

Overall, for demographic variables the average absolute percentage point difference between GOPP and ALLBUS was 4.88 percentage points, between GOPP and ESS: 4.78 percentage points, and 2.05 percentage points between ALLBUS and ESS (not presented in Table 2.3). The difference between the online survey and either of the face-to-face surveys was roughly twice as high as between the two reference surveys.

2.5.2 Attitudinal variables

Among the attitudinal variables, two were present in both reference surveys: self-rated health status and life satisfaction. There was a significant difference in self-rated health between responses from the panel and ALLBUS (8.1 percentage points, p< 0.001). Furthermore, the estimate from ALLBUS is significantly different from the ESS estimate (4.2 percentage point difference, p=0.008).

Life satisfaction was measured with an 11-point scale from 0 “completely dissatisfied” to 10 “completely satisfied.” There was no significant difference in means for ALLBUS and GOPP, however, ESS differed from both GOPP (t=2.83, SE=0.081, p=0.005) and ALLBUS

[14] For χ^2–tests, ALLBUS and ESS distributions are treated as the expected ones, GOPP distribution is treated as observed.

Table 2.3. Comparison of GESIS Online Panel Pilot to two benchmark surveys: demographic and attitudinal variables.

	GOPP			ALLBUS			ESS		
Variable for comparison	**Estimate**	**SE**	**99% Conf. Int.**	**Estimate**	**SE**	**99% Conf. Int.**	**Estimate**	**SE**	**99% Conf. Int.**
Male	51.0%	0.020	[45.8%,56.1%]	51.3%	0.012	[48.2%,54.4%]	53.7%	0.012	[50.7%,56.7%]
Age groups									
18-24	16.6%	0.016	[12.4%,20.8%]	13.9%	0.008	[11.8%,16.0%]	14.8%	0.008	[12.6%,16.9%]
25-34	20.6%	0.017	[16.3%,24.8%]	18.7%	0.009	[16.3%,21.1%]	17.4%	0.009	[15.1%,19.7%]
35-49	34.0%	0.018	[29.2%,38.8%]	35.5%	0.011	[32.5%,38.4%]	33.3%	0.011	[30.5%,36.1%]
50-64	21.2%	0.016	[17.2%,25.3%]	23.0%	0.010	[20.4%,25.6%]	25.7%*	0.010	[23.1%,28.3%]
65+	7.6%	0.010	[5.0%,10.1%]	8.9%	0.007	[7.2%,10.7 %]	8.8%	0.007	[7.1%,10.5%]
Education									
low	12.8%	0.014	[9.1%,16.5%]	23.3%***	0.010	[20.7%,26.0%]	***20.1%******	0.010	[17.7%,22.6%]
medium	30.0%	0.019	[25.1%,34.8%]	35.7%**	0.011	[32.8%,38.7%]	37.9%***	0.011	[35.0%,40.8%]
high	57.2%	0.020	[52.0% ,62.4%]	41.0%***	0.012	[37.9%,44.0%]	42.0%***	0.012	[39.0%,44.9%]
Married	48.5%	0.020	[43.4%,53.7%]	56.9%***	0.012	[53.8%,59.9%]	57.1%***	0.012	[54.1% ,60.1%]
Immigration background	9.6%	0.012	[6.6%,12.6%]	13.4%**	0.008	[11.3%,15.6%]	11.6%	0.008	[9.7%,13.6%]
Working	68.0%	0.019	[63.1%,72.9%]	72.4%*	0.011	[69.7%,75.2%]	***65.8%***	0.011	[63.0%,68.7%]
Good/very good health	60.5%	0.020	[55.4%,65.5%]	68.6%***	0.011	[65.7%,71.4%]	***64.4%***	0.011	[61.5%,67.3%]
Life satisfaction (11-pt)	8.56	0.066	[8.39,8.73]	8.48	0.042	[8.37,8.59]	***8.33*****	0.046	[8.21,8.45]
Trust others									
Yes	19.4%	0.016	[15.3%,23.4%]	23.7%*	0.010	[21.0%,26.3%]	–	–	–
No	19.3%	0.016	[15.2%,23.4%]	35.1%***	0.011	[32.2%,38.1%]	–	–	–
Depends	61.3%	0.019	[56.3%,66.3%]	41.2%***	0.012	[38.2%,44.2%]	–	–	–
German economy (5-pt)	3.39	0.028	[3.32,3.46]	3.09***	0.019	[3.04,3.14]	–	–	–
Self-rated financial situation (5-pt)	3.43	0.030	[3.36,3.51]	3.47	0.020	[3.42,3.52]	–	–	–
German economy in 1 year (5-pt)	2.90	0.032	[2.82,2.98]	3.08***	0.021	[3.02,3.13]	–	–	–
Self-rated financial situation in 1 year (5-pt)	3.20	0.031	[3.12,3.28]	3.18	0.017	[3.14,3.23]	–	–	–
Interest in politics (4-pt)	2.66	0.029	[2.58,2.74]	–	–	–	2.77**	0.019	[2.72,2.82]
Satisfaction with government (11-pt)	4.78	0.087	[4.56,5.01]	–	–	–	4.77	0.051	[4.64,4.90]
Satisfaction with health services (11-pt)	5.38	0.087	[5.16,5.61]	–	–	–	5.81***	0.055	[5.67,5.96]

*p<0.05, **p<0.01, ***p<0.001. Differences between ALLBUS and ESS at the significance level of at least 0.05 in bold italic font; SE short for standard error.

(t=2.42, SE=0.062, p=0.016). Respondents of the panel and ALLBUS demonstrated significantly higher level of life satisfaction than the ESS respondents.

Some concepts were operationalized differently in ESS and ALLBUS: two variables measuring generalized trust and interest in politics, though present in both reference surveys, were not directly comparable. We had to decide which survey had to be used as a reference. The question to measure trust asked whether most people could be trusted, one couldn't be careful enough, or it depended [on a situation]. Respondents in ALLBUS were less trusting than respondents in the panel: the difference on the first value (people can be trusted) of 4.3 percentage points (p=0.022) was substantially lower than differences on categories "one can't be careful enough" (15.8 percentage points, p< 0.001) and "it depends" (20.1 percentage points, p< 0.001) with respondents in the panel having a larger share of those saying trust is situational. The second variable present in both surveys but measured differently was interest in politics. It was replicated from ESS. The analysis showed that respondents in ESS were significantly more interested in politics than respondents in GOPP.

For other attitudinal measures included in the present analysis, only pairwise comparisons with either of the reference surveys could be performed. The next block of measures concerned the economic situation in Germany and respondents' personal financial situations. Online respondents rated the current German economy somewhat better than ALLBUS respondents, and next year's economy worse. No significant differences were found in ratings of respondents' own economic situation for the current and next year.

With respect to ESS, no significant differences were found in satisfaction with government. However, online respondents were less satisfied with the health system (t=-4.19, p<0.001).

2.5.3 Comparison of the GESIS Online Panel Pilot to ALLBUS with post-stratification

From the previous analysis it remains unclear what caused the differences between the surveys. The bias may stem from nonresponse or coverage issues. It is therefore needed to rebalance the GOPP to the Internet-using population. In the absence of official benchmarks or other suitable data, the ALLBUS subsample of Internet users was taken to construct the post-stratification weights. Variables used to construct the weights included age, gender, and educational level. Table 2.4 presents the comparison of weighted estimates (design and post-stratification weights) from the panel to the design weighted ALLBUS data.

Post-stratification did not substantially change the estimates for the demographic variables of marital and employment status, and immigration background: differences between proportions in Table 2.3 and Table 2.4 range between 0.2–1.5 percentage points for these variables. The most pronounced effect was on the variable self-assessed health with 5.6 percentage point difference. Means varied in the range of 0–0.18 scale points. Taken together, post-stratification caused no improvement. On the contrary, with the exception of two categories of trust, status of German economy and self-rated financial situation in one year, for all other variables not used for weighting the differences to the ALLBUS estimates

Table 2.4. Comparison of the GESIS Online Panel Pilot to the subsample of ALLBUS Internet users, with post-stratification.

	GOPP (weighted on ALLBUS Internet users)			ALLBUS Internet users		
Variable for comparison	**Estimate**	**SE**	**99% Conf. Int.**	**Estimate**	**SE**	**99% Conf. Int.**
Married	47.0%	0.025	[40.6%,53.4%]	56.9%***	0.012	[53.8%,59.9%]
Immigration background	8.4%	0.013	[5.1%,11.7%]	13.4%**	0.008	[11.3%,15.6%]
Working	67.8%	0.025	[61.3%,74.2%]	72.4%	0.011	[69.7%,75.2%]
Good/very good health	54.9%	0.025	[48.4%,61.4%]	68.6%***	0.011	[65.7%,71.4%]
Life satisfaction	8.38	0.087	[8.16,8.61]	8.48	0.042	[8.37,8.59]
Trust others						
Yes	17.8%	0.018	[13.1%,22.4%]	23.7%**	0.010	[21.0%,26.3%]
No	23.1%	0.022	[17.3%,28.8%]	35.1%***	0.011	[32.2%,38.1%]
Depends	59.1%	0.025	[52.8%,65.5%]	41.2%***	0.012	[38.1%,44.2%]
German economy	3.28	0.038	[3.18,3.38]	3.09***	0.019	[3.04,3.14]
Self-rated financial situation	3.34	0.037	[3.25,3.44]	3.47**	0.020	[3.42,3.52]
German economy in 1 year	2.85	0.043	[2.74,2.96]	3.08	0.021	[3.02,3.13]
Self-rated financial situation in 1 year	3.20	0.042	[3.09,3.31]	3.18	0.017	[3.14,3.23]

*p<0.05, **p<0.01, ***p<0.001. SE short for standard error. GOPP uses overall weight (design*post-stratification constructed on the basis of ALLBUS Internet users), ALLBUS is design weighted.

increase. Changes in significance can be documented for employment status and self-rated financial situation.

In Table 2.5 we present the comparison between the weighted Online Panel data and the ALLBUS complete sample. Here, the complete ALLBUS sample was used to construct the weights. The results of this exercise are primarily to illustrate the effect of the design decision to exclude non-Internet users. With an exception of life satisfaction and self-rated financial situation, all other variables differ significantly between the two surveys. These differences are largely attributed to the differences between Internet users and nonusers, which are present in ALLBUS for all variables except for marital status (calculations not presented in Table 2.5).

2.5.4 Additional analysis: Regression

It may be hypothesized that discrepancies of attitudinal variables between the surveys may be attributed to differences in sample composition described earlier in this chapter. All demographic variables of the panel with the exception of gender varied from the estimates of the reference surveys by lesser or greater degree. Post-stratification with “rebalancing” the online sample to the Internet using population from ALLBUS did not solve the problem. Another method to assess differences in attitudinal variables is to compare how well the surveys perform in the model estimation. To illustrate this point, life satisfaction, a variable measured equally in all three surveys

Table 2.5. Comparison of the GESIS Online Panel Pilot to ALLBUS, with post-stratification.

	GOPP (weighted on ALLBUS complete)			ALLBUS Internet & Non-Internet users		
Variable for comparison	**Estimate**	**SE**	**99% Conf. Int.**	**Estimate**	**SE**	**99% Conf. Int.**
Married	48.1%	0.036	[38.9%,57.4%]	57.5%*	0.010	[55.1% ,60.0%]
Immigration background	7.8%	0.014	[4.1%,11.5%]	16.7%***	0.007	[14.8%,18.6%]
Working	62.0%	0.038	[52.3%,71.8%]	57.8%	0.010	[55.3%,60.3%]
Good/very good health	48.7%	0.028	[41.5%,55.8%]	56.9%**	0.010	[54.4%,59.4%]
Life satisfaction	8.29	0.101	[8.03,8.55]	8.34	0.037	[8.24,8.43]
Trust						
Yes	15.9%	0.019	[11.0%,20.8%]	22.7%***	0.008	[20.6%,24.8%]
No	26.2%	0.037	[16.7%,35.8%]	42.3%***	0.010	[39.8%,44.8%]
Depends	57.9%	0.036	[48.5%,67.3%]	35.0%***	0.009	[32.6%,37.4%]
German economy	3.27	0.038	[3.18,3.37]	3.04***	0.016	[3.00,3.08]
Self-rated financial situation	3.31	0.044	[3.20,3.43]	3.41*	0.017	[3.37,3.45]
German economy in 1 year	2.82	0.052	[2.68,2.95]	3.01***	0.017	[2.97,3.06]
Self-rated financial situation in 1 year	3.18	0.052	[3.04,3.31]	3.09	0.014	[3.06,3.13]

*p<0.05, **p<0.01, ***p<0.001. SE short for standard error. GOPP uses overall weight (design*post-stratification constructed on the basis of ALLBUS complete sample), ALLBUS is design weighted.

was chosen for the following analysis (Table 2.6). Satisfaction with life is an important indicator for welfare and happiness research and is used widely in international comparative studies.

Headey and Wearing (1992) originally identified seven domains which matter most to well-being: marriage and sex, friendship and leisure, material living standards, work and health. The demographic variables used for comparison to the reference surveys cover at least to some extent five of the domains with the exception of material living standard and friendship and leisure. Further, Headey and Wearing describe the importance of major life events, which we also are unable to include in the analysis. The main purpose, however, being comparison of three surveys for the present chapter, would allow for bias due to omitted variables as long as the list of control variables stays unchanged for each of the surveys in question.

As the goal of the regression analysis was not to interpret substantial outcomes, we concentrate on the part of the model, which accounts for survey control variables. One thing worth noting, however, is the level of R^2, which equals 11.1%. Such a figure is not uncommon for models with socio-demographic variables including age, gender, education, and occupation, which often explain no more than 10% of the variance in individual life-satisfaction (Veenhoven, 1996) in nations like Germany.

Controlling for the sample composition, differences between surveys in life satisfaction do not disappear. To further investigate whether the panel is comparable to the reference

Table 2.6. Linear regression on life satisfaction by reference surveys.

	Standardized β	Standard error
Male	-0.039**	(0.055)
Age group		
18-24	Ref.	
25-34	-0.092***	(0.101)
35-49	-0.168***	(0.103)
50-64	-0.098***	(0.110)
65+	0.036+	(0.129)
Education		
low	Ref.	
middle	0.057*	(0.086)
high	0.142***	(0.082)
Married	0.166***	(0.068)
Working	0.077***	(0.074)
Immigration background	-0.047**	(0.096)
Health (very good/good)	0.237***	(0.062)
ALLBUS	-0.032	(0.079)
ESS	-0.063***	(0.079)
N	4596	
R^2	0.110	

+ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. All surveys combined, weighted, GOPP is reference category.

surveys on the life satisfaction variable, we fitted three regression models, separately for each survey (Table 2.7).

Here, we employ three criteria for comparison of the models' estimates: direction, magnitude, and statistical significance. The regression model, whose results are presented in Table 2.6 suggests that the panel differs significantly from ESS on the life satisfaction variable. The comparison of ALLBUS and ESS models reveals that all estimates have the same direction, however, the magnitude differs (Table 2.7). Additionally, statistical significance (p<0.05) fails to match on three variables: two age group variables and immigration background.

The direction of correlation in GOPP matches these of the reference surveys on all estimates except for the immigration background variable (which most likely is caused by the small number of respondents with immigration background in GOPP) and the oldest age group, probably for the same reason of this group being relatively small. With the exception of the employment status variable (working), variables reaching statistical significance match with the variables reaching statistical significance in ESS. This indicates that as far as modeling of life satisfaction goes, GOPP performs not worse in comparison to

Table 2.7. Regressions on life satisfaction by survey.

	GOPP		ALLBUS		ESS	
	β	SE	β	SE	β	SE
Male	-0.039	(0.132)	-0.028	(0.083)	-0.043^{+}	(0.088)
Age group						
18-24	Ref.		Ref.		Ref.	
25-34	-0.125*	(0.221)	-0.081*	(0.155)	-0.085**	(0.161)
35-49	-0.272***	(0.230)	-0.101*	(0.155)	-0.188***	(0.166)
50-64	-0.246***	(0.268)	-0.023	(0.159)	-0.107**	(0.177)
65+	-0.004	(0.297)	0.089**	(0.204)	0.012	(0.201)
Education						
low	Ref.		Ref.		Ref.	
middle	0.046	(0.251)	0.064^{+}	(0.122)	0.053	(0.137)
high	0.187**	(0.223)	0.113***	(0.119)	0.152***	(0.131)
Married	0.220***	(0.156)	0.134***	(0.101)	0.171***	(0.110)
Working	0.054	(0.175)	0.093**	(0.121)	0.076**	(0.111)
Immigration background	0.016	(0.229)	-0.066**	(0.131)	-0.049^{+}	(0.160)
Health (very good/good)	0.172***	(0.140)	0.245***	(0.096)	0.249***	(0.095)
N	746		1831		2019	
R^2	0.128		0.104		0.120	

$^{+}$ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$; data are weighted, beta coefficients are standardized, SE is short for standard error.

ESS than ESS does to ALLBUS or ALLBUS does to ESS. The R^2 of the GOPP model also is not far off from the ESS model's one. This suggests that the data, which comes from GOPP, could be used to fit models (on life satisfaction at the least) and would provide adequate results. The shortcoming of GOPP with the immigration background or oldest age group estimates having an opposite direction would not be relevant if one wished to interpret only significant coefficients, as in ESS both coefficients of immigration background and age group older than 65 do not reach statistical significance of 0.05 or lower.

The results presented in Tables 2.6 and 2.7 may be dependent on the choice of one particular variable: life satisfaction. Therefore, we fitted another model using another variable present in all of the reference surveys: self-assessed health status. As mentioned in section 2.4.2, self-assessed health was recoded into a binary variable "very good/good health". The results of the logistic regression are presented in Table A2 in Appendix A. Controlling for demographics, significant differences between GOPP and both reference surveys are found.

It cannot be inferred from the above whether GOPP will prove to be successful in comparing other variables or not. This would have to be decided for other variables or models case by case, which goes beyond the scope of this chapter.

2.5.5 Replication with all observations with missing values dropped

Pairwise comparisons between surveys, results of which are presented in Table 2.3, were replicated when all observations were dropped, which contained a missing value on at least one of the variables used in the analysis. The resulting estimates did not differ substantially from the results in Table 2.3. The highest difference was 1.3 percentage points for proportions of demographic variables (mean difference 0.42; 0.54 and 0.48 percentage points for GOPP, ALLBUS and ESS respectively). The highest difference for variable means was 0.06 scale points. No changes of significance, except for ALLBUS and ESS immigration background ($p<0.05$), took place. Although levels of significance changed for life satisfaction in GOPP and interest in politics in ESS.

2.6 Discussion and conclusion

In some countries and research settings probability-based online panels have established themselves as a cost-effective alternative to face-to-face and telephone general population surveys. Still, in other countries such interviewing methods are at the beginning of their emergence. In this chapter we described an online panel pilot study whose goal was to identify and test the conditions under which a probability-based panel can be recruited and maintained in Germany. The target population was restricted to the Internet-using population and recruitment was performed by telephone using an RDD dual-frame approach.

Response metrics in the recruitment process reached satisfactory levels, beginning with moderate response rates for the telephone interviews. Due to the experimental nature of various recruitment approaches, the average recruitment rate was lower than one would expect with the optimal design approach.

A major focus of this chapter has been on assessing the quality of the data generated by the panel. Quality was defined in terms of deviation from estimates produced by two major face-to-face surveys, the German General Social Survey (“ALLBUS”) and the German subsample of the European Social Survey (ESS).

The results indicate that significant differences exist between the online panel and either of the face-to-face reference surveys on most of the demographic and attitudinal variables. Out of five demographic measures used for comparison, only one was not significantly different from either of the face-to-face surveys. Most attitudinal variables differed when compared to the reference surveys. Though for a number of attitudinal measures comparisons had to be restricted to one reference survey due to a lack of data. This was the case when concepts were measured differently between the two reference surveys or the concept was only present in one of them.

When estimating data quality we also need to take into account that the face-to-face reference surveys also showed significant differences between each other. Such differences were found for two demographic and on both attitudinal measures. Not only did these two surveys differ on means or proportions, but also on magnitude and significance of predictors when performing a regression analysis. Taken together, the results suggest that the panel is comparable to traditional surveys in many aspects of sample composition and when modeling social phenomena. Taking into account the estimation of average absolute error calculated for demographics of about 2 percentage points between the two surveys and about 5 percentage points between the GOPP and either face-to-face survey, we can conclude that the GOPP estimates still are not as close to the face-to-face survey as one would wish. However, at the present stage of research, when probability-based online panels are still at their beginnings, the results presented are rather encouraging.

This research has a number of limitations. First, a design decision to exclude non-Internet users made comparisons with official statistics impossible as German Census does not include a question on Internet use. This forced us to use other general population surveys as benchmarks restricting these surveys to the subsample of Internet-using population and use one of the surveys for the calculation of post-stratification weights. Using survey data as benchmarks was problematic because both reference surveys had considerable nonresponse. The response rates were not atypical for the current survey climate in Germany, this does not, however, exclude the possibility of bias.

Absence of official statistics is not a problem unique for the group of Internet users. It may apply, for example, to special (hard-to-reach) population groups. One important direction for future research is to employ strategies, which would evaluate bias in reference surveys.

Second, estimates from the online panel are the product of multiple selection processes, each of which is prone to nonresponse: the telephone interview, providing the email address, and response to the first online questionnaire. The mode switch between contacting the respondent and data collection online was an additional source of nonresponse, which neither of our reference surveys had. Respondents, who had agreed to the initial recruitment interview, may have used the time lag between the telephone interview and the invitation to the online survey as an opt-out option. More detailed studies on the causes of these processes and how they influence the differences in estimates would be beneficial.

Third, there is a risk of potential mode effects on measurement. The telephone interview was not only aimed to recruit participants into the panel, but also to collect data about the respondents to study nonresponse effects. In such scenarios mode effects are to be expected (de Leeuw, 2005). It is also worth mentioning that the reference surveys were interviewer-administered whereas the GOPP was self-administered in the online surveys. For comparison we did not use questions, which were considered to produce socially desirable answers and for which interviewer-administrated and Internet modes have demonstrated bias (Chang & Krosnick, 2009; Kreuter, Presser, & Tourangeau, 2008), but there is still a possibility of mode effects on other items due to visual and aural administration of the questions.

Finally, the results of the overall data quality assessment are dependent on the choice of the set of variables to be examined. The approach of assessing the data quality, which we undertook, of comparing means and reporting absolute errors with respective significance tests, is performed in a number of studies which examine the quality of online panels. However, apart from demographic variables, covariates considered by researchers, range broadly from voting behavior and attitudes to immigration to smoking behavior. The inherent problem of performing such multiple comparisons with various measures is the concern about finding more differences when more variables are added to the set. It might be useful to find an appropriate number of measures to be examined.

The limitations of the approach we undertook for the quality assessment of data produced by the online panel, nevertheless, cannot undermine our overall conclusion that the data and performing model-based analysis with this data are comparable to those of the traditional surveys. In our comparisons we concentrate mostly on biases, the magnitude and the direction thereof. In doing this, one needs to take into account that estimates are a product of multiple processes. As Biemer (1988) points out, "a sample survey is an integrated system of activities and operations for data collection" (p. 273), so the comparison takes place between several systems of data collection. We try to take this into consideration by comparing not only final estimates but how surveys perform when modeling social phenomena. However, this strategy also deals with final "survey products". As we show, consequences of certain design decisions cannot be solved with post-stratification either. In future research this should deserve special attention. As the GESIS Online Panel Pilot, data from which we compared to the general population surveys, is a methodological project, in our opinion, it served the purpose well. For conducting general population surveys in order to assess their representativeness, one would want for the design to mirror the procedures implemented in the reference surveys: comparable recruitment mode, comparable experimental conditions and in the best case scenario inclusion of the offline population. For Germany, all these strategies are implemented in the GESIS Panel, a mixed-mode omnibus access panel survey of the general population, which includes both Internet and non-Internet population.

Chapter 3

Nonresponse in the recruitment process of an online panel

3

Abstract

A major concern in probability-based online panels is nonresponse. Due to the absent sampling frame such as a complete list of email addresses for the target population, recruitment of probability-based online panels has to rely on alternative modes of data collection. In addition to noncontact or nonresponse associated with this initial stage, reached respondents may not be willing to join an online panel or, if they are willing to join, may still fail to respond to the online questionnaire. The central questions of this chapter are whether those who agree to participate are different from those who do not, and whether those who participate online are different from those who were willing to participate but failed to do so. We hypothesize that different factors influence the decision to participate in an online panel and the actual participation. In this chapter, we study the role of respondents' demographic and attitudinal characteristics as well as the influence of survey design features, especially the online mode of administration. For this purpose, we examine how respondents' experience with the Internet and their experience with online surveys influence (non)response. We find that demographic characteristics of the respondent, survey design features, and the respondent characteristics specific to the survey mode predict willingness to participate and actual participation in the panel. Some of the demographic differences do not persist after controlling for experience with online surveys and Internet-use related characteristics. We conclude that including respondent characteristics specific to the survey mode benefits the understanding of nonresponse in online panel surveys.

3.1 Introduction

Probability-based online panels are seen as a solution to coverage and selection problems introduced by volunteer online panels, which are much criticized by the scientific community (Baker et al., 2010; Yeager et al., 2011). Online panels that employ probability-based selection of respondents have the advantage of relatively low cost of data collection compared to face-to-face and telephone modes. This is one of the reasons for the recent emergence of online panels, which differ in their recruitment protocols but share the trait of being based on a probability sample. Examples of such panels are the LISS Panel and the CentERpanel in the Netherlands, the German Internet Panel (GIP), the Knowledge Panel of GfK Custom Research (formerly Knowledge Networks), the Gallup Panel, and the RAND panel in the United States.

Nevertheless, questions about the quality of data in such panels continue to be raised. One of the major concerns is nonresponse: falling response rates are documented across modes and across countries (de Leeuw & de Heer, 2002). This problem is more critical for online surveys, because in online surveys response rates are lower than in other survey modes (Lozar Manfreda, Bosnjak, Berzelak, Haas, & Vehovar, 2008). Despite the advantages offered by the online data collection, the key question is: Are high costs of recruitment and maintenance of probability-based online panels justified given the high nonresponse? Nonresponse does not necessary equal nonresponse bias (Groves & Peytcheva, 2008), but since response rates are viewed as one of the indicators of survey quality, this question deserves further investigation.

This chapter focuses on nonresponse in the recruitment process for an offline-recruited probability-based online panel of Internet users in Germany. The goal is to understand the selection mechanisms, which contribute to the nonresponse structure of the multistep recruitment procedure. Nonresponse mechanisms known from the literature are complemented by studying respondent characteristics specific to the online surveying mode.

3.2 Background

A substantial body of research on nonresponse in online surveys has accumulated over the years. It should be noted that nonresponse in online surveys needs to be studied with an eye on the recruitment procedures that are used to gain participants' agreement to participate in an online survey. Two groups of recruitment procedures for online surveys can be distinguished: "simple" recruitment, where participants can be directly invited to take part in an online survey, and "complex" recruitment, where multistep procedures of contact and invitation are employed. Using the typology of online surveys (Couper, 2000), intercept surveys and list-based surveys of high-coverage populations, which target special groups such as web site visitors or university students, fall into the group of online surveys using a simple recruitment procedure. Mixed-mode surveys with a choice of a completion mode, pre-recruited panels of Internet users, and online surveys based on probability samples of the entire population fall under a complex recruitment procedure.

In such complex procedures, a random sample is drawn (either from population registers, an RDD-sample, or an address-based sample) and potential respondents are contacted via telephone or face-to-face with or without postal prenotification and asked to participate in an online panel. Some organizations provide Internet access and the necessary equipment for conducting online surveys to those respondents who do not have it (for example, the LISS panel, Knowledge Panel and GIP – the German Internet Panel); still others do not do so and restrict their target population to Internet users. Figure 3.1 depicts a typical recruitment process for a probability-based online survey of the Internet-using population with an initial interview carried out via telephone. Individuals contacted in this manner have known non-zero selection probabilities, which can be used for the calculation of response rates, weights, and post-survey adjustments. Moreover, information about all respondents from the recruitment interview can be used to study whether the resulting sample is biased in respect to the characteristics of interest. One major problem of such recruitment protocols is the absence of information on nonrespondents to the recruitment interview. This problem is solved in cases in which register or frame data is available on both respondents and nonrespondents. However, the number and/or the informativeness of available characteristics is usually limited.

Sample representation has been a focus of several studies that employed a multistep recruitment procedure (e.g., Bartsch, 2012; Bosnjak et al., 2013; Hoogendoorn & Daalmans, 2009; Lee, 2006). Generally, recruitment procedures in these studies share several stages, each of which is susceptible to some kind of loss of potential respondents: (1) initial contact and agreement to participate in an initial (recruitment) interview, (2) intention to participate in the online survey or online panel, (3) actual participation in the online survey. In some cases, an additional stage is present: selection from the pool of respondents who agreed to participate (following stage 2). However, that stage is carried out by the researcher and therefore does not contribute to the nonresponse process.

For prerecruited online surveys of Internet users, it has been shown that the most pronounced differences in sample composition are found between those who have Internet access and those who do not (Bandilla, Kaczmirek, Blohm, & Neubarth, 2009; Couper, Kapteyn, Schonlau, & Winter, 2007). Couper et al. (2007) found significant differences in socio-economic status and health-related variables between those who had Internet access

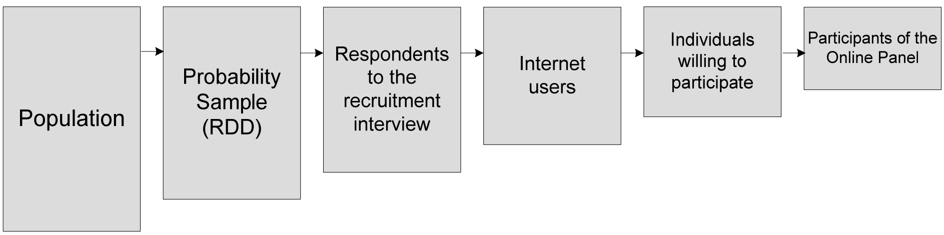


Figure 3.1. The recruitment process of an online panel of Internet users with an initial telephone interview.

and those who did not for the sample of individuals over 50 years of age. Bandilla et al. (2009) showed demographic differences between Internet users and the full sample primarily in respect to age and education. At the stage of expressing willingness to participate in an online survey, given the respondent had Internet access, some differences persisted. Couper et al. (2007) showed that significant predictors of willingness to participate in an online survey were age, race/ethnicity, and education. Bandilla et al. (2009) found that younger respondents and higher educated respondents were more likely to be willing to participate in an online survey. However, the studies concluded that nonresponse at the final stage was not a major concern when compared to noncoverage and to the initial nonresponse: Once willingness to participate in the online survey had been expressed, only marginal differences remained between those who agreed to participate and those who actually started an online survey. There is a reason to be optimistic about noncoverage: a study by Mohorko, De Leeuw, and Hox (2013) shows that coverage bias is diminishing over time and across countries. Nonresponse, nevertheless, remains problematic.

The studies by Bandilla et al. (2009) and Couper et al. (2007) asked respondents to the initial interview to participate in a one-time online survey. The studies whose goal was to recruit respondents into an online panel and where those not having access to the Internet were provided with it, thus eliminating the coverage error, also reported significant nonresponse effects. In every study, demographic characteristics were found to differ between the initial sample and the respondents to the online interview. Lee (2006) found that those willing to participate in the Knowledge Panel were more likely to be unemployed, to live in larger households, and to be higher educated. She reported some differences in ethnic composition as well. For the stage of actual participation these differences, however, were practically nonexistent. Similar findings were reported by Couper et al. (2007), Bandilla et al. (2009), and Bosnjak et al. (2013). Hoogendoorn and Daalmans (2009) found age and income to be predictive of nonresponse in all recruitment stages for the CentERpanel with older individuals and the recipients of lower income being less cooperative. Little selectivity was found with respect to the ethnic composition. For the LISS Panel, selectivity with respect to household size, age, education, marital status, gender, urbanicity, and immigration background of a certain generation was found (Scherpenzeel & Das, 2011).

Researchers mostly concentrate on studying demographic differences between respondents and nonrespondents and between individuals willing to take part in the online panel and those unwilling to do so. Demographic characteristics are usually readily available but are not usually expected to explain nonresponse; rather they serve as evidence of the underlying mechanisms. One such mechanism is the familiarity with the medium used to conduct the survey. There is evidence, that despite the provision of the Internet access and equipment, computer ownership is predictive of willingness to participate and of final response in online panels. Lee (2006) found prior web experience to be predictive of the willingness to participate on a statistically significant level. For the LISS panel, Internet access prior to the provision of the equipment was reported to be predictive of response to the online panel (Scherpenzeel & Das, 2011). Couper et al. (2007) hypothesized that experience with and being comfortable using the Internet was a possible explanation of nonresponse, but they were unable to test this hypothesis with their data.

Other potential determinants of expressing willingness and final response lie in the design features of the survey. Couper et al. (2007) found the length of the initial recruitment interview to be positively correlated with the willingness to complete the Internet survey. This seemingly contradictory finding has been documented for face-to-face panel surveys: The length of the interview was positively associated with response in the subsequent wave (Zabel, 1998). The potential explanation for this is that the length of the survey is not indicative of burden in that case but rather of respondents' engagement with the survey. Another design feature affecting response is the use of incentives. It is well known that incentives enhance response probability in self-administered postal and web surveys as well as interviewer-mediated surveys (see the reviews by Singer, 2002 and Singer & Ye, 2013). In her literature review and a meta-analysis on prepaid and promised incentives, Göritz (2006) showed that incentives positively influenced response in list-based and nonprobability online surveys. Little yet is known about the effects of incentives in probability-based online panels. Scherpenzeel and Toepoel (2012) found that prepaid incentives increased the registration rate by about 15 percentage points compared to a no-incentive group.

Even if not explicitly stated, these studies follow the Groves and Couper (1998) framework for survey participation in household interview surveys. The framework uses characteristics of the social environment, respondents' characteristics, survey design features, interviewers' characteristics, and interaction with an interviewer to explain the survey response process. In the studies on recruitment for online panels, the interviewer's characteristics and interaction with an interviewer have not been given as much attention as other characteristics. Hoogendoorn and Daalmans (2009) included paradata about CATI-contact attempts in their analysis, but these were not specific to the recruitment interview for an online panel. One notable exception to include the interaction with interviewers during the recruitment process is a dissertation by Vehre (2011), in which she analyzed various strategies interviewers used to convince respondents to participate in an online panel depending on the reasons for initially expressed refusal to participate. Vehre (2011) found that arguments tailored to the reasons for unwillingness to participate in an online panel had the highest potential to convert initially reluctant respondents to join the panel. This finding is in line with Groves, Cialdini, and Couper (1992), who showed that tailoring by interviewers is crucial for survey response.

The standard approach to study nonresponse during the recruitment process is to rely on several explanations of willingness to participate using respondent characteristics such as demographics, Internet usage patterns, and paradata. One possible problem is that this approach may not take into consideration the nature of the multistep nonresponse process, where different factors may become relevant at different stages. One way of addressing this issue is to use different theoretical frameworks for various recruitment stages. One example is employing the framework based on the theory of planned behavior (Ajzen, 1991) for analyzing actual survey participation. This framework has been successfully applied to studying nonresponse in online surveys (see e.g., Bosnjak, 2003; Heerwegh, 2005; Vehre, 2011).

A model based on the theory of planned behavior assumes that respondents make a definite decision to participate during the recruitment interview and then follow their decision. However, when multiple decisions have to be made before the actual participation, providing an email address (an indicator of the definite decision) may not necessarily equal the actual intention to participate. Apart from making a definite decision to participate during the recruitment interview, two scenarios are possible. First, respondents to the recruitment interview may be unsure whether they want to participate or not and leave the option open to be invited into the panel in order to postpone the decision. Second, respondents to the recruitment interview may decide during the interview that they do not want to participate but still provide an email address because they do not want to be impolite. In the course of the project GESIS Online Panel Pilot, which will be described later, the invitation email was sent briefly after the respondent had provided the email address at the end of the recruitment interview. Some email addresses proved to be invalid ("bounced"). Respondents to whom an email invitation could not be delivered were contacted afterwards with a request to correct the email address and some respondents refused to do so. This may be regarded as a hint that individuals provided the wrong email addresses to avoid participation but expressed willingness in order to maintain the trustful atmosphere of the recruitment interview. On the other hand, a number of respondents may decide not to participate in online surveys during the recruitment interview but provide the correct address not to be confronted with possible further questions about reasons for noncooperation.

For these reasons, the actual intention cannot be assumed equal to providing the email address. Therefore, it seems best to use the Groves and Couper (1998) framework, which differentiates between the factors influencing the survey cooperation decision as being outside of the researcher's control (exogenous) and those under the control of the researcher (endogenous). There are two groups of characteristics which are beyond the researcher's control: characteristics of the social environment (e.g., survey-taking climate, economic situation, and the neighborhood characteristics) and characteristics of the sampled person/household (e.g., social characteristics, psychological attitudes, structure of the household). The two blocks including characteristics which the researcher can control are survey design features (e.g., survey mode, respondent selection procedure, topic of the survey) and the characteristics of interviewers (e.g., experience, socio-demographic characteristics).

Dillman, Eltinge, Groves, and Little (2002, p. 6) showed how causes of nonresponse differed by modes of data collection. Thus, the framework to study nonresponse in online panels is adjusted to include Internet-related factors (Figure 3.2). The complex nature of the decision process for the final response (i.e., agreement to join and starting participation in an online panel) is captured by leverage saliency theory, which acknowledges the "threshold nature of the survey response propensity" (Groves, Singer, & Corning, 2000, p. 300). According to leverage saliency theory, each sampled individual attaches weights to various attributes of the survey, which are highlighted (made salient) during the survey request by an interviewer or in an advance letter. One potential respondent may attach more leverage to a particular survey attribute whereas another potential respondent may not value the particular survey attribute that highly. Depending on which attributes are made salient during the survey request and what leverages potential respondents attach to these attributes, the decision is made whether to cooperate with the survey request or refuse to participate. In the context of a multistep recruitment, leverage saliency theory should capture that different factors' varying levels of importance salience across the various recruitment stages. For instance, interaction with an interviewer should be more influential during the recruitment interview and less important for the actual online participation. The framework also allows for leverages of different attributes to vary across respondents. Thus, the online mode of administration may be viewed as a negative attribute by respondents who have less familiarity with the mode and as a positive attribute by respondents who feel more comfortable with the Internet and online surveys.

In summary, I expect various respondent characteristics and survey design attributes to have different levels of leverage across respondents and different levels of salience across multiple recruitment stages.

3.3 Mechanisms and hypotheses

Social environmental factors

Social environmental factors are typically exogenous to the survey design (Dillman et al., 2002) and include the characteristics of the context in which a survey request is made (Groves & Couper, 1998, p. 31). The role of social environmental factors in explaining nonresponse to survey requests can be characterized by the growing number of unsolicited interview requests and telephone calls by telemarketers. Such calls raise hostility towards survey requests in the United States (Tourangeau, 2004) and in Europe (for the Netherlands see de Leeuw & Hox, 2004; for Germany, see Vehre, 2011). Concerns about unsolicited sales

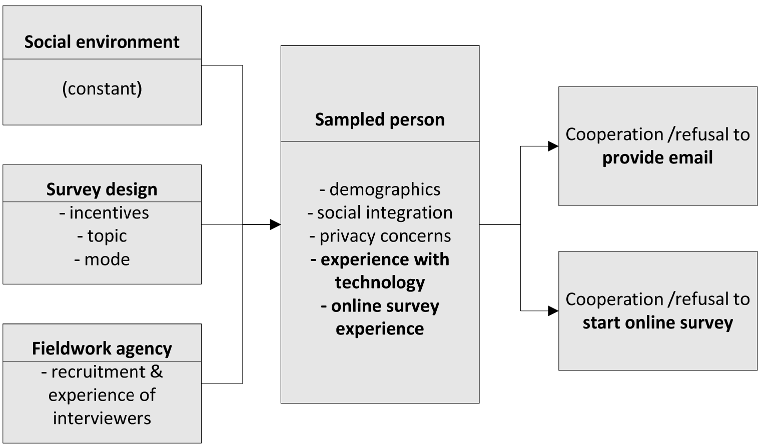


Figure 3.2. Framework to study nonresponse in the course of the recruitment process for an online panel.

propositions occur at the initial contact stage. We expect that these concerns mostly affect the response to the recruitment interview and will not affect the willingness to participate in an online panel and actual participation.

Respondent characteristics

Respondent characteristics form the second block of relevant nonresponse factors exogenous to survey design. They include the socio-economic characteristics and psychological dispositions that influence the participation decision. Groves and Couper (1998) argue that while socio-economic characteristics are not causal for participation decisions, they influence psychological predispositions towards surveys. These predispositions may include topic salience, reactions to strangers (fear of crime) and other feelings (e.g., helpfulness or embarrassment), which influence respondent-interviewer interactions. Explanations of how socio-psychological characteristics of the respondent influence the decision to participate in a survey are provided by social exchange theory and by social isolation theory. Social exchange theory as applied to survey response behavior states that a respondent as a rational actor reciprocates the efforts of the research organization (Dillman, 1978) or of various societal institutions (Goyder, 1987). Social isolation is a related explanation that explains nonresponse as a behavior of individuals who are not guided by the mainstream culture, who feel isolated from a wider society, and who are not guided by a norm of "civic duty" that would prompt them to support surveys carried out by the central institutions of society and hence do not comply with the survey request (Groves & Couper, 1998, p. 131). Persons who refuse to participate in an interview but are then persuaded to take part in a follow-up survey, not only have been found to be less socially active but to be less involved in classical culture, sports, and reading (Stoop, 2005, p. 237). Groves and Couper (1998) further state that the decision to participate in a survey is situational rather than thought-through in advance. This decision is based on heuristics that apply specifically to the survey-request situation. Thus, respondent characteristics which are not specific to the survey participation mode are expected to have a powerful influence at the stage of agreeing to the initial recruitment interview and not to contribute much at the following willingness and online participation stages.

Survey design features

The first block of factors influencing the survey participation decision under the control of the researcher are survey design features. These include incentives, interviewer training which teaches interviewers how to tailor the persuasion strategies, the information provided in an advance letter, and introductions made by interviewers. This block of characteristics also includes burden: interview length and topics perceived as threatening (Groves & Couper, 1998, p. 36). One aspect of survey design features, which has received much scholarly attention and will be examined here, is the use of incentives. Incentives are expected to have a positive effect on willingness to join the panel. The positive effect of incentives should also hold for the step of actual participation since there are no theoretical reasons for incentives to become irrelevant at this stage. If a respondent has a high leverage for the incentive at the willingness stage and the decision to agree to participate in the panel is based on the attractive prospect of an incentive, this effect should manifest itself at the step of actual participation as well.

Interviewer characteristics and interaction with an interviewer

The final block of factors which are under the control of the researcher are the characteristics of the interviewer, which include socio-demographic and professional attributes (Groves & Couper, 1998). According to leverage saliency theory, interaction of the potential respondent with an interviewer plays a crucial role in making various attributes of a survey salient for the respondent. If the interviewer has convinced the respondent to participate in an initial interview, the interviewer can also be expected to have an influence on a potential respondent's willingness to join an online panel. However, it is not expected that the interaction with an interviewer is still influential at the stage of the actual participation. Interaction with an interviewer during the initial interview may be viewed as a characteristic of the situation in which the decision is made (Groves et al., 1992). The situation, in which the final decision to participate is made, is characterized by the absence of an interviewer, therefore, the interviewer's effect is not expected to be found at the stage of the actual participation.

Respondent characteristics specific to the survey mode

The extension of the framework for survey cooperation includes respondent characteristics, which relate to the respondent's predispositions to online surveys and may influence the response decision. Online surveys may be new for certain groups of respondents, in which case respondents' experience with this medium may be one of the factors influencing the final decision to participate in an online panel. In line with the findings of Lee (2006), Hoogendoorn and Daalmans (2009), and Scherpenzeel and Das (2011), prior experience with the Internet is expected to have an influence on both willingness to participate and on actual participation. Nonresponse specific to the online mode may be driven by the computer literacy and by perceived ability to manage filling out the online questionnaire (Bosnjak, 2003). Thus, even after expressing willingness to participate, respondents who are less experienced with the Internet and email may not attempt to click on a link or register for the online survey at a website. Once the influence of experience with technology is controlled for, there is no theoretical reason to expect demographic characteristics such as age and education (which may be indicative of the ability to use the Internet) to be predictive of nonresponse.

Another respondent characteristic specific to the survey mode is experience with online surveys. There is some empirical evidence of the influence on how prior experience with surveys influences individual willingness to participate in an online survey. Couper et al. (2007) found that reluctance to take part in the initial interview led to unwillingness to participate in an online survey, Bartsch (2012) showed that having refused a survey request in the past led to a lower willingness to join an online panel. These findings concern prior survey experience irrespectively of the mode of administration. Here it is hypothesized that having prior experience with online surveys increases the willingness to participate in an online panel. This effect is expected to hold for the step of actual participation as well.

3.4 Data and methods

The data for the analyses were collected in a pilot project "GESIS Online Panel Pilot"[15] carried out by GESIS – Leibniz Institute for the Social Sciences in Mannheim, Germany. Following a short pretest, GESIS started the telephone recruitment for a probability-based online panel in February 2011. Due to the decrease of landline coverage in Germany (Gabler, Häder, Lehnhoff, & Mardian, 2012), a dual-frame approach was chosen to perform the recruitment with 50% landline and 50% cell phone numbers in the frame. To recruit the participants, the randomized last digit method was used which is a variation of a random digit dialing (RDD) for Germany (Gabler & Häder, 2002). Up to 15 call attempts were made by the interviewers to reach the sampled number. For the landline cases, a respondent within a household was chosen via the last birthday method. No such strategy was necessary for the cell phone cases, since sharing a cell phone device in Germany is assessed to be about 2% (Gabler et al., 2012) and therefore a person answering a cell phone was automatically considered to be the respondent.

The target population was limited to German-speaking adults aged 18 and older, who lived in Germany and used the Internet for non-work-related purposes. GESIS did not provide the equipment for those without an Internet access. A screening question about non-work-related Internet use was asked during the telephone interview, and those without Internet access or using the Internet for work-related purposes only, did not proceed to the follow-up questions of the interview.

The standard interview duration was about 10 minutes. The interview contained questions on demographics, some attitudinal questions, and questions on telephone use (for weighting purposes). At the end of the interview, a recruitment question was asked prompting the respondent to provide their email address in order to be invited to join an online panel. The panel was described as monthly questionnaires on various topics (family life, work, free time activities, health) that take 10–15 minutes to complete, experimentally assigned incentive amounts were promised for the completion of every online questionnaire excluding the control group, which received no payment. An option to view the project's or the institute's web site was offered to reluctant respondents, followed by a telephone call with a second recruitment attempt.

One of the goals of the project was to optimize the recruitment process. Therefore, the recruitment was performed in three sequential parts referred to below as Study 1, Study 2, and Study 3. There were important variations across these study parts (see the overview in Table 3.1). The variations important for the analysis presented here concerned the fieldwork period, the incentive amount and the fieldwork agency, which performed the recruitment. The fieldwork for Study 1 started in February 2011 and ended in April 2011. The incentives of 5 or 10 Euros were promised for each completed monthly questionnaire with an additional 20 Euros bonus after eight months if no questionnaire was skipped. The fieldwork for Study 2 was conducted from June 2011 to July 2011. The incentive groups were changed to 5 Euros, 2 Euros and a control group receiving no payment. No bonus was offered. The fieldwork for

15 https://dbk.gesis.org/dbksearch/sdesc2.asp?no=5582&db=e&doi=10.4232/1.11570

Studies 1 and 2 was carried out internally by GESIS using its in-house telephone interviewing facilities. The fieldwork for Study 3 began in June 2011 and ended in August 2011. The incentive conditions were identical to those of the Study 2. Study 3 was carried out by an external fieldwork agency with interviewers having some experience in conducting studies for academic research but primarily having experience in conducting market research interviews.

Overall, 95279 telephone numbers comprised the initial sample. The overall response rate for the telephone interview (AAPOR 3) across all studies was 17.8%. The number of completed telephone interviews across studies added up to 4840 interviews. Among these, 3514 respondents reported use of the Internet for non-work related purposes (72.6%). The survey invitation was sent to 1665 respondents who provided a valid email address. The first online survey was started by 1010 persons and completed by 934 persons.

There were experimental variations in the course of the panel recruitment, which also are described in Table 3.1. The first variation across all studies was a split-ballot experiment, in which half of the respondents were asked the categorized income question and half were not. The second variation was offering a respondent a short version of the questionnaire, which was used when multiple call appointments failed to reach the respondents and for those respondents refusing telephone interviews in principle. Also, a small proportion of respondents was assigned to short interviews for experimental purposes. All short interviews are excluded from further analysis because they only contain a question on Internet use and few demographic questions, not allowing further inference. Excluding the short interviews decreased the sample size to 2631 Internet users for whom information from the telephone

Table 3.1. Overview of the features of the GESIS Online Panel Pilot.

Characteristic **Fieldwork start**	**Study 1** **February 2011**	**Study 2** **June 2011**	**Study 3** **July 2011**
Fieldwork agency	In-house, own telephone lab	In-house, own telephone lab	External, specialized on market research surveys
Internet-screener	The very first question with target sample unit	Preceded by a general attitudinal question	Preceded by a general attitudinal question
Incentive experiment	- 10 EUR plus 20 EUR bonus for completion of all 8 questionnaires - 5 EUR plus 20 EUR	- 5 EUR - 2 EUR - 0 EUR	- 5 EUR - 2 EUR - 0 EUR
Announced panel duration	8 months	no limitation	no limitation
Length of recruitment interview	- regular version	- regular version (90%) - shorter version (10%)	- regular version (90%) - shorter version (10%)
Experiment on household income question	- household income question asked in 50% of cases	- household income question asked in 50% of cases	- household income question asked in 50% of cases
Sample size, telephone numbers	20155	17547	57577
Response to the telephone interview (AAPOR RR3)	23%	26%	14%

recruitment interview is available. Of these 2631 respondents, 1356 individuals provided their email address to be invited to the online panel and 789 persons started the first online survey.

To analyze nonresponse at the various recruitment stages, multivariate logistic regression models were used, which had two dependent variables: (1) willingness to join an online panel and (2) response to the first online questionnaire. The first binary-coded variable “willing to participate” contrasts those who completed the full telephone interview and provided their email address and those who refused to provide the email address. The rare cases when a respondent stated not having an email address (N=75) can be classified either as ineligible or as refusing to join the online panel since the possibility to register with an email provider in order to take part in the survey is open to respondents. For the purposes of the analyses presented here, respondents stating they had no email address at the end of the recruitment interview were coded as non-willing to join the panel.

The second outcome variable “online response” refers to those individuals, who started answering the survey questions online versus those who did not login to the survey or did not proceed to the survey questions. Only those who actually started filling in the survey were treated as respondents since those who log in and quit after seeing the first screen do not provide any data and in this respect are equal to unit nonrespondents. On the other hand, those who log into the online survey may be more similar to respondents than to those who never log in to the survey. The reasons for starting and for not continuing the survey may be explained with a different set of mechanisms. Peytchev (2009) argues that respondents who start the survey and then break it off are different from unit nonrespondents. Initial unit nonrespondents do not see the characteristics of the web page and of the online questionnaire. In our case, there is reason to believe that the disclaimer statement, which was presented on the first page, might have influenced the response behavior causing the break-off at the disclaimer page. Therefore, further analyses primarily concern those who started the online survey, also treating those individuals who viewed only the first screen as respondents.

Previous research has shown the importance of distinguishing between nonresponse due to noncontact and nonresponse due to refusal (Groves & Couper, 1998; Lynn & Clarke, 2002; Steele & Durrant, 2011). In our study, respondents who agreed to participate in the panel were invited per email. If the invitation email could not be delivered (“bounces”), we were faced with nonresponse due to noncontact if the email was valid, that is, the email address existed but the invitation could not be delivered due to spam-filters or because the postbox was full. If the email was not valid, that is, no such email address existed, the case might be classified as noncontact or as an implicit refusal to join the panel. Two scenarios how to treat the bouncing cases were possible. First, owners of those emails could be classified as willing to join an online panel (hence, providing the email address) but ineligible due to noncontact. Second, owners of such email addresses could have been unwilling to join an online panel and provided a false address in order to finish the telephone interview. Respondents, whose survey invitations could not be delivered, were contacted by interviewers with a request to correct the wrong address, so if the invitation stayed undeliverable, it may be hypothesized that inexistent email addresses belonged to the respondents not willing to join the panel.

However, since not all owners of bounced emails could be reached by telephone to correct the email address, it was likely to be a combination of both unwillingness and noncontact. After checking the bouncer status of the provided email addresses, 47 cases which had the status "bounces again after correcting the email address" or "household/respondent not available for call-back" could have been dropped. However, 4 of these cases started the online survey. We therefore take a conservative approach and do not differentiate between noncontact and noncooperation, treating all nonresponse as noncooperation.

In the remainder of this section, the nonresponse mechanisms described in section 3.3 are operationalized. All mechanisms with the exception of social environment are studied in this chapter. Social environmental factors are outside the control of the researcher, as has been noted earlier. In the telephone interview mode, among the social environmental factors, one of the major concerns is the use of the interview requests by telemarketers. As Singer and Presser point out, "although interviewers may be able to counter concerns about privacy and confidentiality if they are raised by respondents in face-to-face interviews, the tendency to hang up in the first few seconds of a phone call from a stranger often precludes this strategy in telephone surveys" (2008, p. 470). In addition to this characteristic, the RDD sampling strategy provided no frame information that could have been used as a proxy to study the social environmental factors. The focus of this chapter therefore is on endogenous nonresponse factors under the control of the researcher and those exogenous factors we can observe.

The respondents' characteristics included primary demographics such as age, gender, and education (three groups based on a variable measuring secondary education status), and demographic characteristics intended to measure social connectedness/social integration such as immigration background, partnership status (married or living together with a partner as a reference), having children under 18 years of age in the household. The direct measure of social integration included generalized trust, a question which had three answer options: one can trust people, it depends, and one cannot be careful enough (reference category). In order to capture the amount of discretionary time, employment status (coded as working or nonworking for students, people who stay at home, and people in retirement), and direct measures of being generally active such as helping neighbors/relatives, going out, and doing sports were asked. In order to avoid triggering distrust during the telephone interview, direct questions on privacy concerns were not asked.

Survey design features included incentive conditions and the fieldwork agency. Incentives were incorporated into the analysis as four groups: 0, 2, 5, and 10 Euros. The variable fieldwork agency took the value 0 for GESIS and 1 for an external fieldwork agency. This variable included various aspects of differences in interviewer's training, selection, and experience. The difference in experience of interviewers conducting research primarily for academic versus for market research purposes is thought to influence the tailoring strategies for persuading respondents to join the online panel. Interviewers conducting research primarily for academic purposes (i.e., GESIS in-house interviewers) are trained to persuade the respondent to take part in the telephone interview once they established contact. Although it might not be true for all institutions whose studies are primarily nonacademic, some of the fieldwork agencies concentrate on the number of completed interviews rather than persuading a reluctant respondent to take part in the interview. The striking difference in telephone response rates reported in Table 3.1 between Studies 2 and 3 with equal incentive conditions supports this explanation. However, it is not assumed that all individual differences between the interviewers that may influence their recruitment strategies are captured by the fieldwork variable.

The last mechanism was represented by respondent characteristics specific to the online survey mode. The first group of characteristics represented experience with technology, using the following indicators: duration of Internet use (measured in years), frequency of Internet use, and whether the respondent was interviewed on his or her cell phone versus being interviewed on a landline telephone. The last indicator was meant to capture affinity towards (new) technology. This indicator was not perfect for representing the concept since being interviewed by landline or by cell phone was the result of being part of the landline or the cell phone component of the frame. An alternative indicator, available from the recruitment interview, was the ownership of a cell phone for those interviewed via landline. However, this indicator did not measure actual use of the device. For this reason, the former indicator was used. The last respondent characteristic specific to the online mode was previous experience with online surveys. During the interview, it was not specified what we understood under the term “online surveys,” it thus encompassed anything from occasional polls on websites to academic surveys. Online surveys were not defined explicitly in order to include everything the respondent considered an online survey, since the main goal was to capture familiarity with online interviewing as perceived by the respondent.

3.5 Results

3.5.1 Willingness to join the online panel

In order to predict willingness to participate in the online panel, three logistic regression models were fit to the data (Table 3.2). The first model included respondent demographic characteristics and indicators of social integration and activity (Model 1). The second model additionally included survey design features. The third model added respondent characteristics specific to the online surveying mode (Model 3).

Model 1 shows that willingness to join the online panel is higher among younger individuals, men, and higher educated individuals. These differences mirror some of the well-known demographic disparities between the Internet users and nonusers (Initiative D21 & TNS Infratest, 2013; Lenhardt et al., 2003; NTIA - National Telecommuniation and Information Administration & Economics and Statistics Administration, 2013; Wagner, Pischner, & Haisken-DeNew, 2005) despite the fact that only Internet users were interviewed by telephone. Further, married respondents and respondents living with a partner are less willing to join the online panel. Respondents working for pay also showed significantly lower willingness to join the online panel. Having children aged less than 18 years in the household does not predict willingness to participate at a statistically significant level.

3

Table 3.2. Logistic regressions with the dependent variable “willingness to join the online panel.”

	Model 1		Model 2		Model 3	
Variable	**Odds ratio**	**SE**	**Odds ratio**	**SE**	**Odds ratio**	**SE**
Demographics						
Age	0.984**	(0.003)	0.983**	(0.003)	0.989**	(0.004)
Gender (male)	1.274**	(0.110)	1.287**	(0.113)	1.144	(0.105)
Education (low)	ref.	—	ref.	—	ref.	—
middle	1.283+	(0.166)	1.267+	(0.167)	1.155	(0.158)
high	1.470**	(0.186)	1.474**	(0.189)	1.171	(0.160)
Working (yes)	0.761**	(0.074)	0.773**	(0.077)	0.808*	(0.083)
Immigrant (yes)	0.826	(0.103)	0.841	(0.106)	0.897	(0.117)
Married/partner	0.741**	(0.071)	0.726**	(0.071)	0.706**	(0.072)
Children <18	1.010	(0.100)	0.989	(0.100)	1.014	(0.106)
Social Integration						
Going out	0.987	(0.050)	0.962	(0.049)	0.933	(0.050)
Sport	1.014	(0.038)	1.007	(0.038)	0.995	(0.039)
Helping out	1.001	(0.053)	1.007	(0.054)	1.034	(0.057)
Trust (no)	ref.	—	ref.	—	ref.	—
yes	1.489**	(0.205)	1.473**	(0.206)	1.347*	(0.195)
it depends	1.341**	(0.138)	1.351**	(0.141)	1.322**	(0.143)
Survey Design						
Incentive (none)			ref.	—	ref.	—
2 Euros			1.102	(0.130)	1.095	(0.134)
5 Euros			1.458**	(0.163)	1.480**	(0.171)
10 Euros			2.571**	(0.446)	2.742**	(0.494)
Fieldwork agency			0.772**	(0.073)	0.740**	(0.073)
Survey-specific characteristics						
Interviewed on cell					1.240*	(0.115)
Internet use, years					1.016+	(0.009)
Frequency of Internet use					1.386**	(0.080)
Online survey experience					2.082**	(0.243)
Constant	1.754+	(0.566)	1.799+	(0.621)	0.275**	(0.123)
Sample size	2340		2340		2300	
Pseudo R^2	0.02		0.05		0.08	
BIC	3255.930		3218.352		3086.623	

Note: + $p<0.10$, * $p<0.05$, ** $p<0.01$, unweighted data, SE is short for standard error.

Predictors which aimed at measuring social integration are for the most part not statistically significant. The only significant predictors are levels of generalized trust: those respondents saying others are generally trustworthy or saying "it depends," are significantly more likely to be willing to join the online panel compared to those saying people cannot be trusted.

Adding the survey design characteristics to the model does not substantially change the odds ratios. Incentives do not have an overall positive effect on the willingness to join an online panel. There seems to be no statistically significant difference between offering no incentives and 2 Euros, however offering 5 or 10 Euros per survey significantly increases respondents' willingness to participate in the online panel. Along with the incentives, the fieldwork agency plays an important role in the recruitment process. The commercial fieldwork agency, specializing in non-academic studies with interviewers whose primarily expertise is in conducting market research surveys, has significantly lower chances to recruit participants into the panel than does the group of in-house interviewers trained in conducting academic surveys.

Model 3 additionally controls for respondent characteristics specific to the online mode of administration. The covariates related to the Internet experience were expected to be significant predictors of willingness. Experience with the Internet, however, has only a marginally statistically significant association with the willingness to join the online panel. On the other hand, respondents, who use the Internet more often, are more willing to join the panel. When the covariate "frequency of Internet use" is not controlled for, the predictor "Internet use in years" reaches statistical significance (OR=1.028, $p<0.01$, SE=0.009; Table B2 in Appendix B).

Online survey experience shows the highest odds ratio of 2.082 among mode-specific respondent characteristics. Respondents who have completed one or more online surveys in the past are more likely to be willing to join the panel. This effect is highly significant and is in line with Bartsch's finding (2012, p. 116) that those with online survey experience were more likely to agree to further participate in a follow-up online survey after a recruitment interview. In sum, all online-survey-specific characteristics are related to nonresponse at the willingness stage. Those respondents who have more experience with the Internet and online surveys are more likely to express willingness to join the panel. Moreover, the demographic gender and education variables no longer show statistically significant effects and their odds ratios decrease once online-related covariates are controlled for. Contrary to expectations, the effect of age does not disappear. The effect of being married or living with a partner is also significant across all models. Trust levels are also significant predictors of willingness after controlling for mode-specific factors.

3.5.2 Response to the online questionnaire

The second dependent variable was the actual response to the online questionnaire. The results of the logistic regressions predicting the response to the first online survey are presented in Table 3.3. Respondent characteristics are examined in Model 4. It was expected that variables capturing the mechanism of social integration would not be relevant predictors of response to the online survey. No significant effects are found for the measures

3

Table 3.3. Logistic regressions with the dependent variable "response to the first online questionnaire."

	Model 4		Model 5		Model 6	
Predictor	**Odds ratio**	**S.E.**	**Odds ratio**	**S.E.**	**Odds Ratio**	**S.E.**
Demographics						
Age	1.014**	(0.005)	1.013**	(0.005)	1.013*	(0.005)
Gender (male)	1.016	(0.125)	1.042	(0.130)	0.990	(0.128)
Education (low)	ref.	—	ref.	—	ref.	—
middle	1.095	(0.210)	1.108	(0.217)	0.907	(0.185)
high	1.798**	(0.337)	1.827**	(0.350)	1.371	(0.277)
Working (yes)	1.023	(0.135)	1.026	(0.138)	1.044	(0.146)
Immigrant (yes)	0.455**	(0.081)	0.484**	(0.087)	0.458**	(0.085)
Married/partner	0.982	(0.132)	0.935	(0.128)	0.910	(0.128)
Children <18	1.039	(0.146)	1.050	(0.150)	1.096	(0.162)
Social Integration						
Going out	0.964	(0.069)	0.940	(0.069)	0.913	(0.069)
Sport	0.974	(0.052)	0.955	(0.052)	0.930	(0.052)
Helping out	0.941	(0.071)	0.958	(0.074)	0.981	(0.078)
Trust (no)	ref.	—	ref.	—	ref.	—
yes	1.446+	(0.285)	1.424+	(0.287)	1.361	(0.282)
it depends	1.274	(0.192)	1.250	(0.192)	1.228	(0.194)
Survey Design						
Incentive (none)			ref.	—	ref.	—
2 Euros			1.325	(0.231)	1.395+	(0.251)
5 Euros			1.868**	(0.307)	1.914**	(0.322)
10 Euros			2.546**	(0.392)	2.729**	(0.640)
Fieldwork agency			0.734*	(0.101)	0.732*	(0.103)
Survey specific characteristics						
Interviewed on cell					0.878	(0.112)
Internet use, years					1.042**	(0.014)
Frequency of Internet use					1.296**	(0.118)
Online survey experience					1.564**	(0.228)
Constant	0.686	(0.309)	0.577	(0.280)	0.158**	(0.102)
Sample size	1233		1233		1215	
Pseudo R2	0.04		0.06		0.09	
BIC	1716.089		1699.613		1666.570	

Note: significance levels + p<0.10, * p<0.05, ** p<0.01, unweighted data.

of social integration. However, immigration background, which may also be seen as an indicator of social integration, becomes relevant at a stage of the response to the first wave of the online panel. Respondents with immigration background who agreed to participate in the online panel are significantly less likely to start the online survey. The measures of generalized trust are no longer significant at this stage.

Among the demographic characteristics, gender no longer shows significant effects for online response, but there is a curious effect of age – the coefficient changes direction and remains statistically significant. One possible reason for this is that once committed to the survey, older respondents are more likely to execute their choice and start the survey. The same explanation may apply for reversed sign of the coefficient for employment status: if respondents think they will not have time for participation in the panel, they do not agree to participate (at the telephone stage). Once they have agreed to participate, no differences are found between working and nonworking respondents. However, the effect of employment status is not significant. The effect of incentives does not substantially change between the models (Model 2 vs. Model 5). However, contrary to expectations, the fieldwork agency effect still holds at the stage of the response to the first online survey.

The full model for the online response (Model 6) shows that online survey specific covariates (experience with the Internet, frequency of Internet use, and experience with online surveys) predict the response to the first wave of the panel. Recruitment interview mode (cell phone versus landline) does not influence the response to the first online survey.

After adding the online survey specific covariates to the model, some changes in parameter estimates for demographic characteristics occur (Model 4 compared to Model 5 and Model 6). As was the case with age at the stage of willingness to participate, the effect does not disappear after controlling for Internet- and online survey related predictors. Having the highest level of education, however, no longer predicts the response to the online questionnaire at a statistically significant level. Nonresponse associated with immigration background is not explained by an additional set of predictors, which is supportive of the hypothesized explanation of weaker identification with German society and a possible perception of nonspecific general population surveys as not being beneficial for the part of the population with immigration background.

Lastly, online survey experience has a reduced odds ratio when compared to the full model of willingness to participate (Model 3, Table 3.2). Although the models cannot be directly compared with each other since they are co-dependent (only those who agreed to participate included into the participation models), one may speculate that having experience with online surveys has more leverage for respondents at the stage of agreeing to participate than for actual online response. As was the case with the willingness models, model fit is low: McFadden R^2 ranges from 0.04 to 0.09. However, the Bayesian Information Criterion (BIC) shows no penalty for adding the covariates, indicating the importance of including survey design features and survey mode-specific characteristics.

3

3.6 Discussion

This chapter studied nonresponse at two stages of the recruitment process common for probability-based online panels. The models tested the influence of respondent socio-demographic characteristics and survey design features known to affect survey response. The models were further extended by adding respondent characteristics specific to the online mode. Nonresponse mechanisms were tested across two recruitment stages: (1) expressing the willingness to join an online panel by providing an email address and (2) response to the first wave of the online panel.

The results showed that nonresponse was selective with respect to respondent characteristics at both stages. However, different sets of variables predicted the response at each stage. First, we studied the influence of respondents' characteristics. Among demographic characteristics, age, gender, education, employment status, and marital status were significant predictors of nonresponse at the willingness stage. A possible explanation why married respondents or respondents living with a partner were less willing to join the online panel could be the absence of discretionary time for the completion of several online questionnaires. This also would explain the highly significant lower willingness to join the online panel by respondents working for pay.

It was hypothesized that individuals more integrated into society would be more open to participation in surveys and therefore more willing to join the online panel. Among attitudinal variables measuring whether respondents were socially active and trusting, only generalized trust influenced the willingness to participate. Respondents saying "others are generally trustworthy" or saying "it depends" were significantly more likely to be willing to join the online panel compared to those saying "people cannot be trusted." These covariates may partially have captured the effect of having privacy concerns on giving away one's email address. Privacy concerns with online surveys might be relevant to the willingness to participate and the subsequent response, but they could not be measured directly during the telephone interview in order to avoid the risk of lowering the willingness to participate in the panel by making the privacy concerns salient to the respondent. Although the association between generalized trust and privacy concerns is not straightforward, they have been shown to be related. Couper, Singer, Conrad, and Groves (2008) found that trust (along with sensitivity and attitudes toward privacy and surveys) significantly predicted respondents' perceptions of harm. Joinson, Paine, Buchanan, and Reips (2007) showed that trust and privacy both moderated self-disclosure, to which the researchers attributed why respondents may have renounced their privacy concerns when a survey request came from a trusted organization.

At the subsequent stage of online participation, age, education, and immigration background significantly predicted response to the online survey. Immigration background was not a significant predictor of willingness to participate. Any explanations of this finding can only be speculative, because they could not be tested explicitly. Since individuals with immigration background had completed the telephone interview, a lower response propensity cannot be explained by insufficient language skills. The fact that individuals with immigration background did not differ significantly from those having no immigration background at the willingness stage may indicate that respondents with an immigration background agreed to the telephone interview and agreed to participate in the online panel because they did not want to be impolite during the conversation with an interviewer. However, at the stage of actual online participation, when they did not have to worry about being perceived as impolite, they became noncooperative.

Second, the influence of survey design features was studied. These features included incentives and the fieldwork agency. The fieldwork agencies performing the recruitment differed on a variety of characteristics. The first fieldwork agency was an in-house telephone lab with interviewers experienced primarily in academic surveys and with a spatial proximity to the researchers. The second fieldwork agency was external with interviewers specializing primarily in market research surveys. Both incentive conditions and fieldwork agency were predictive of response at both recruitment stages, contrary to the expectation that the influence of the fieldwork agency would cease at the online response stage. This might suggest that interaction with interviewers produced differential level of commitment, which was reflected at the stage of online response.

Third, we added some mode-specific respondent characteristics such as experience with Internet technology, being interviewed on a cell phone, and experience with online surveys to explain the willingness and subsequent (non)response. Being interviewed on a cell phone was a significant predictor of willingness to join the panel but was no longer significant and the coefficient changed direction at the online response stage. Frequency of Internet use and having experience with online surveys were both significant predictors of willingness and of actual online response. The chances of providing an email address at the recruitment interview were twice as high for those experienced with online surveys compared to those having no experience with online surveys. For starting the online survey, this effect decreased somewhat but nevertheless stayed significant. These additional technology- and online-survey-related variables could explain some of the demographic differences at the willingness stage: gender and education were no longer significant predictors of willingness when the technology-related variables were added to the model. Gender was also no longer significant in the online response stage once the technology characteristics were added. However, those online-mode-specific characteristics could not fully explain the differences in demographics. For example, employment status stayed predictive of response at the willingness stage and immigration background stayed predictive of response at the online response stage after controlling for mode-specific covariates. The finding that nonresponse associated with immigration background was not explained by an additional set of predictors supports the weaker-identification-with-society explanation offered above.

One limitation of this study was that although it captured characteristics of the respondent, which were relevant to the fact that the panel was an online panel as opposed to the initial telephone interview, the specific attitudes towards switching the interviewing mode from telephone to online could not be accounted for. It has been shown that a mode-switch can be a cause of nonresponse (Fricker, Galesic, Tourangeau, & Yan, 2005; Kreuter, Presser, & Tourangeau, 2008; Sakshaug & Kreuter, 2011). Thus, it would be beneficial to disentangle the mode change from other nonresponse mechanisms.

Another limitation was that no frame information had been available on nonrespondents to the telephone interview. Thus, the nonresponse process studied in this chapter did not account for selection processes, which took place at the stage of agreeing to take part in a telephone interview. We assumed that societal level predictors of nonresponse and mechanisms of social integration played a role at this initial stage, but this could not be investigated with the available data. Lastly, the model fit (pseudo-R^2) was very low for the basic models: 0.02 and 0.04. Adding survey design features and specific respondent characteristics led to increases although the overall fit stayed low (0.08 and 0.09).

Nevertheless, the Bayesian Information Criterion (BIC) informative for comparisons of non-nested models, decreased when design features and characteristics specific to the mode were added to the models. This means that complementing nonresponse mechanisms with characteristics related to prior Internet and online survey experience seems to benefit our understanding of (non)response at the selection steps of initial willingness expression as well as response to the first online survey.

Chapter 4

Respondent attrition: The role of incentives and survey experience

This chapter was co-written with
Lars Kaczmirek and Edith de Leeuw.

Abstract

One of the major concerns about data quality in panel surveys is the loss of panel members (i.e., panel attrition). In addition to reduced sample size, attrition may introduce bias to the estimates if nonresponse is selective. To minimize the negative consequences of panel attrition – either by preventing attrition through survey design or by correcting the consequences of attrition with weighting – it is important to study the reasons behind it. In this chapter, we apply the benefit-cost theory of survey participation to studying panel attrition. This theory emphasizes intrinsic motivations for survey participation. We focus on respondents' survey experience (intrinsic) and offered incentives (extrinsic) as motivational factors. We seek to answer the question whether incentives inhibit panel attrition and if so whether they serve as a compensation for a negative experience or serve as an additional benefit. We further differentiate between several attrition groups based on their response patterns. We find that both incentives and survey experience affect panel attrition. Incentives do not compensate negative panel experiences, rather they serve as an additional bonus to positive survey experience. The distinctive feature of our analyses is the focus on time-varying covariates of survey experience. Based on our findings, we suggest that time-varying factors should be given more attention when explaining panel attrition.

4.1 Introduction

Panel surveys offer several advantages over cross-sectional surveys in terms of data analysis and data collection (e.g., Cantor, 1989; Lynn, 2009, p. 388). However, panel surveys have unique sources of error which do not exist in cross-sectional surveys. The major nonsampling error concern in panel surveys is attrition. Attrition refers to nonresponse specific to panel surveys, when respondents discontinue participation after having completed the first round or several single wave studies of the panel. The loss of respondents over time is problematic for two reasons. First, it decreases the precision of estimates due to the reduced sample size. More importantly, it increases the risk of bias in the survey estimates if those respondents who stay with the panel systematically differ from nonrespondents (attriters).

In order to reduce attrition, it is important to understand the mechanisms responsible for the loss of panel members. Most studies fall short on providing theoretical considerations about the nature of attrition, focusing on correlates of attrition instead (e.g., Behr, Bellgardt, & Rendtel, 2005; Burkam & Lee, 1998; Hawkes & Plewis, 2006; Tortora, 2009; Watson, 2003; Zabel, 1998). The focus on correlates of attrition serves the goals of assessing the generalizability of the results obtained through the panel study and constructing the weights for the correction of attrition bias rather than explaining attrition. Lepkowski and Couper (2002) propose a framework to study attrition as an outcome of multiple processes. They distinguish between three distinct processes that lead to attrition in (household) surveys: location, contact given location, and cooperation given contact. These processes are driven by different theoretical predictors, which fall into two major groups: survey design features and respondent characteristics. Cooperation propensity is also predicted by household-interviewer interaction. Apart from the location component, this framework resembles the framework for studying nonresponse in cross-sectional surveys: household-level nonresponse is also subject to differentiating between noncontact and noncooperation (Groves & Couper, 1998).

According to Kalton, Kasprzyk, and McMillen (1989), the reasons for nonresponse in panel surveys are the same as the reasons for nonresponse in the initial wave of the panel (ibid., p. 250). There is, however, reason to believe that panel nonresponse is different from cross-sectional nonresponse. The major difference is that panel respondents have completed one wave of the survey and know what is expected of them in future surveys. The decision to participate in later waves of the panel survey is more considered and less "automatic" than the decision of one-time participation (Schnell, 1997, p. 143). Several studies investigate the mechanisms of refusals in panel surveys differentiating them from the mechanisms of nonresponse in cross-sectional surveys. Prominent mechanisms of nonresponse in panel surveys are: panel fatigue, absence of commitment, absence of participation habit, and the influence of life events. The first mechanism, panel fatigue, occurs when respondents feel that they have "done enough," become bored and uninterested in the survey and, ultimately, nonrespondents (Laurie, Smith, & Scott, 1999; Lugtig, 2014). In the second mechanism, panel attrition is explained by the fact that members lack commitment to participate in a panel survey in the first place (Lemay, 2009; Lugtig, 2014). The third mechanism is the absence of participation habit. Making the decision to continue participation is costly for panel members both in terms of time and effort required to integrate new information. Therefore, panel members follow their prior decisions: participants develop the habit of participation whereas attriters do not (Lemay, 2009). Lastly, panel members may become nonrespondents as a result of life events such as changes in living arrangements, financial, or employment situation (Short & McArthur, 1986; Trappmann, Gramlich, & Mosthaf, 2012; Van den Berg, Lindeboom, & Dolton, 2006), and similar situational characteristics, such as physical or mental health, economic situation, financial stress experienced by respondent (Lepkowski & Couper, 2002), and major life events or "shocks" (Lemay, 2009) such as the birth of a child, death of a household member etc.

Although these mechanisms acknowledge that attrition is different from initial nonresponse, they cannot be viewed as a single framework. They operate on different levels, mixing the characteristics of the survey (panel fatigue), the situation (life events), the respondent (absence of commitment), and the decision process (habit). Several studies find that previous participation patterns are highly predictive of future response behavior: participation in the previous wave or percent of previously completed interviews predict future cooperation (Olsen, 2005; Watson & Wooden, 2009), and previous irregular participation or the growing number of waves of successive nonparticipation predict attrition (Das, 2012; Watson & Wooden, 2011). Nevertheless, previous response history does not provide insight into the reasons for attrition or response in panel surveys.

In this chapter, we propose to study attrition within the benefit-cost theory of survey participation (Singer, 2011). According to this theory, people participate in surveys when, in their subjective judgment, the benefits of participation outweigh the costs (ibid., p. 380). This theory resembles other theories of survey participation: social exchange theory (Dillman, 1978), leverage-saliency theory (Groves, Singer, & Corning, 2000), and the theory of planned behavior and reasoned action (Ajzen, 1991) in its application to survey participation (Bosnjak, 2003; Hox, De Leeuw, & Vorst, 1995). The benefit-cost theory of survey participation differs from social exchange theory in emphasizing the costs of participation, and it differs from leverage-saliency theory in that the cost-benefit theory does not focus the role of the interviewer in making the factors relevant for participation salient to the respondent. Benefit-cost theory can be seen as a "synthesis of principles derived from other theories" (Singer, 2011, p. 388). Our proposal to use this theory for studying panel attrition is not unique. Former studies have employed similar rational-choice-based frameworks to understand attrition. Hill and Willis (2001) applied economic modeling based on social exchange theory in which a respondent makes a decision to participate in a subsequent wave if expected utility of the interview process plus incentive payment exceed the expected time costs of the interview. Fitzgerald, Gottschalk, and Moffitt (1998) argue that a respondent compares the value of participating to the value of not participating, Lugtig (2014) draws on the leverage-saliency theory of survey participation to explain panel attrition. In contrast to these theories, the benefit-cost theory of survey participation emphasizes the role of the intrinsic motives of participation (Singer, 2011).

The participation decision for the next wave in the panel is made at a lower level of uncertainty and the information about the survey acquired in the first panel wave can be either attributed to benefits or to the costs of participation depending on the experience the respondent had in the initial wave of the panel survey. The experience with the initial and subsequent panel waves relates to the content of questionnaires, incentive payments, and, in interviewer-administered panels, the influence of interviewers. Pleasant survey experience is a benefit which may encourage respondents to participate in the future waves. Watson and Wooden (2009) show that respondents' perceptions of the interview experience is the single most important predictor of cooperation in the future survey waves. Using factor analysis, Hill and Willis (2001) also demonstrate that two factors – engagement and ease of understanding the questionnaire – are the most powerful and significant predictors of the further participation in a panel.

Negative experience within a panel adds to the costs and inhibits further participation. For instance, interviewer observations of respondents being hostile, restless, or impatient during the previous interview negatively affect cooperation in future waves (Olsen, 2005). Another indicator of unpleasant interview experience is item nonresponse; respondents showing a high degree of item nonresponse are more likely to refuse future interviews (Burkam & Lee, 1998; De Keulenaer, 2005; Hawkes & Plewis, 2006; Lee, Hu, & Toh, 2004; Loosveldt, Pickery, & Billiet, 2002; Watson & Wooden, 2009). Length of the interview, as an indicator of perceived respondent burden (Sharp & Frankel, 1983), shows unexpected results for panel surveys: Some studies find a positive relationship of interview length with attrition (Zabel, 1998), in other studies, longer interviews are positively related to the propensity to take part in future interviews (Branden, Gritz, & Pergamit, 1995; Hill & Willis, 2001; Olson & Witt, 2011; Watson & Wooden, 2009). The positive association between interview length and subsequent response could reflect commitment to the panel survey.

The use of (monetary) incentives is another survey-related feature often conceptualized as a benefit of survey participation. Incentives have been shown to increase response rates in all survey modes in cross-sectional studies (see meta-analyses of Church, 1993; Göritz, 2006; Singer, Van Hoewyk, Gebler, Raghunathan, & McGonagle, 1999), and in longitudinal studies (e.g., Castiglioni, Pforr, & Krieger, 2008; Göritz, 2008; Laurie & Lynn, 2009; Scherpenzeel & Toepoel, 2012; Singer & Ye, 2013). The results of incentive manipulations, however, are inconclusive. Introducing an incentive has a positive effect of cementing the loyalty of respondents (Jäckle & Lynn, 2008; Laurie, 2007), increasing the incentive over the course of a panel shows a significant increase of response rates of cooperative sample members (Laurie, 2007; Laurie & Lynn, 2009; Rodgers, 2011). However, experimental evidence also shows that incentives appear to work only on a minority of respondents (Olsen, 2005). Furthermore, ceasing to provide or decreasing the incentive amount does not have a significant effect on panel participation (Lengacher, Sullivan, Couper, & Groves, 1995; Singer, Groves, & Corning, 1999). In an incentive experiment, Trussell and Lavrakas (2004) find that cooperating respondents are more likely to subsequently participate even with no incentive offered than noncooperative respondents who were offered various incentive amounts. Zagorsky and Rhoton (2008) report similar results for a mature panel. Taken together, these findings may indicate that incentives have an additional value when a respondent makes the decision to participate in the next wave or not. However, for respondents who stay with the panel for a longer period of time there are other (intrinsic) benefits involved which overweigh the costs of participation, even when the incentives are taken out of the equation.

The goal of this chapter is to study the benefits and costs that influence the decision to continue participation or to become a nonrespondent in a panel survey. We do so by contrasting survey experience and incentives. This draws on the distinction made within the benefit-cost theory of survey participation between intrinsic and extrinsic motivational factors. Singer (2011) broadly defines intrinsic motivation in terms of formal properties of the survey (appearance of the questionnaire) and mode of administration (pleasure of talking with an interviewer). In our definition, specific to panel survey participation decisions, intrinsic motivation includes pleasant experience in the previous wave and extrinsic motivation is provided by incentives. Both factors are attributed to the benefits of longitudinal participation. The costs of longitudinal participation include the longitudinal burden. In an experimental study, Apodaca, Lea, and Edwards (1998) find that higher commitment is needed from respondents when they are asked to join the panel than when they are not informed about the subsequent interviews. The authors differentiate between an immediate burden – the length of the initial interview – and longitudinal burden – the perceived cost of continued participation. We expand this definition of longitudinal burden to include the accumulated burden over the course of the panel. Longitudinal burden includes the aspects of response burden as defined by Bradburn (1978): interview length, respondent effort, respondent stress, and frequency of being interviewed. With every panel wave, respondents reevaluate their initial decision to participate in the panel survey. If participation benefits to the self or to society outweigh participation costs, respondents continue participation. Otherwise, panel members decide to skip the wave or discontinue participation altogether. The benefits of participation may be intrinsic or extrinsic. The costs of participation include several aspects of respondent burden.

The following hypotheses aim at uncovering the relationship between intrinsic and extrinsic panel survey benefits and costs.

Hypothesis 1: Positive experience in the panel inhibits attrition, whereas negative (burdensome) experience increases the likelihood of nonparticipation;

Hypothesis 2: Incentives decrease the likelihood of attrition;

Hypothesis 3: When the experience of the previous waves in the panel was positive, incentive amounts do not influence the likelihood of further participation/attrition;

Hypothesis 4: When the experience of the previous waves within the panel was burdensome, incentives serve as a compensation for this negative experience;

Hypothesis 5: Respondents motivated by positive experience and respondents motivated by incentive payments have dissimilar participation patterns.

As we have noted earlier, the majority of studies on panel attrition investigate the role of respondents' socio-demographic characteristics. It has been shown that older panelists are more likely to attrite (Behr et al., 2005; Branden et al., 1995; Burkam & Lee, 1998; Fitzgerald et al., 1998; Peracchi, 2002; Watson & Wooden, 2009), some studies have found that the attrition likelihood is also high among young panel members (Behr et al., 2005; Peracchi, 2002; Tortora, 2009; Watson & Wooden, 2009). Men were found to attrite at a higher rate than women (Behr et al., 2005; Burkam & Lee, 1998; Fitzgerald et al., 1998; Lepkowski & Couper, 2002; Lynn, Buck, Burton, Jäckle, & Laurie, 2005; Peracchi, 2002; Watson & Wooden, 2009). Respondents with immigration background or who belong to minorities are more likely to attrite (Branden et al., 1995; Burkam & Lee, 1998; Fitzgerald et al., 1998; Lynn et al., 2005; Tortora, 2009; Watson & Wooden, 2009), and attrition is lower among highly educated respondents (Behr et al., 2005; Fitzgerald et al., 1998; Lepkowski & Couper, 2002; Lynn et al., 2005; Peracchi, 2002; Tortora, 2009; Watson & Wooden, 2009) and respondents with higher income (Behr et al., 2005; Fitzgerald et al., 1998; Lynn et al., 2005; Watson & Wooden, 2009). Generally, people with more stable living situation are less likely to attrite. In line with this pattern, attrition likelihood is lower among those who are married (Behr et al., 2005; Fitzgerald et al., 1998) and those who have children (Branden et al., 1995). Conversely, attrition likelihood is higher among those living alone (Behr et al., 2005; Watson, 2003), not married (Peracchi, 2002), unemployed (Behr et al., 2005), living in urban areas (Branden et al., 1995; Lynn et al., 2005; Watson & Wooden, 2009), and with a greater likelihood of moving (Behr et al., 2005; Watson, 2003).

Although some of these characteristics may be only relevant for face-to-face surveys since they relate to probabilities of locating panel members, other characteristics may be related to attrition irrespective of survey mode. For example, higher attrition among men in face-to-face administered surveys persists after controlling for contact (Lepkowski & Couper, 2002), and for an online-administered Gallup Panel men were more likely to attrite as well (Tortora, 2009). Thus, we include these socio-demographic characteristics into our analyses. However, respondents' socio-demographic characteristics are not hypothesized to influence attrition directly. They rather indicate differences which exist in the benefit-cost calculation processes of different respondent groups.

4.2 Data and method

4.2.1 Data

The data were collected in 2011 and 2012 in a probability-based panel of Internet users in Germany, the GESIS Online Panel Pilot (GOPP), which was a methodological project at GESIS – Leibniz-Institute for the Social Sciences. Respondents were recruited via telephone interviewing. A dual frame approach, which included landline and mobile numbers, was used (for more information on recruitment of the panel, see Struminskaya, Kaczmirek, Schaurer, & Bandilla, 2014). The target population consisted of German-speaking adults (18 years and older) residing in private households in Germany, who used the Internet for non-work-related purposes. Within the household, the target person was chosen using the last birthday method. Individuals who agreed to participate in the telephone interview and who used the Internet for non-work-related purposes were asked to complete an interview of varying length. The length of the interview was varied experimentally with a 10 minutes (short) or 15 minutes (normal) duration. The normal-length interview included attitudinal questions, demographic questions, and questions on telephone usage for the weighting purposes. The short version included one attitudinal question and the demographic questions. At the end of the interview, respondents were asked to provide their email addresses in order to join the online panel. In several cases interviewers were allowed to skip the interview questions and proceed directly to the recruitment. These conditions included respondents opposing telephone interviews in general, hard refusals due to "too busy," and unavailability of the respondent for at least three appointments made with the interviewer.

Respondents who agreed to participate in the online panel were sent an email invitation every month for a survey with a duration of 10–15 minutes. The overall response rate (AAPOR RR3) for the telephone interviews was 17.8%. Overall, 4840 telephone interviews with an Internet usage rate of 72.6% (3514 respondents) were conducted. Among these, 1665 respondents agreed to participate in the online panel and provided a valid email address to receive an invitation to the online survey. From those, 1010 respondents started the first online survey and 934 persons completed it. The recruitment rate (the proportion of respondents who provided initial consent over all eligible respondents) was 9%.

Several design features are relevant for studying attrition. The online panel started with the first introductory multi-topic questionnaire, whose goal was to raise respondents' interest in the panel and encourage further participation. Starting with wave two, each monthly questionnaire had a main topic to which most of the questions were devoted. Questionnaire 2 (Q2) concentrated on education and employment, further topics were: family life (Q3), religion and values (Q4), environment and ecology (Q5), social networks (Q6), politics (Q7), and a multi-topic questionnaire focusing on gender roles and personality (Q8). Although the panel continued after eight waves, after wave 8 respondents were explicitly asked whether they wanted to continue participation. Further design features such as a bonus payment after eight waves and an incentive reduction after wave 8 make attrition in the later waves incomparable to the first eight waves of the panel. Therefore, we restrict the analysis of attrition to the first eight waves of the GOPP.

Recruitment for the panel was done in three sequential steps and due to the experiment on panel conditioning (see Chapter 5), the order of questionnaires Q5 and Q6 was changed for some respondents (Table 4.1). Furthermore, respondents from samples 1 and 2 were recruited by interviewers of an in-house GESIS telephone surveying center, whereas respondents from samples 3a and 3b were recruited by an external agency, specializing primarily in market research. A short nonresponse follow-up study was conducted for sample 2. Nonrespondents converted to respondents as a result of this action are excluded from the attrition analysis, because resulting from their participation timing and additional recruitment effort, these participants are not comparable to the regular panelists.

Table 4.1. Questionnaires within the GESIS Online Panel Pilot

Date (approx.)	Mar 2011	Apr 2011	May 2011	Jun 2011	Jul 2011	Aug 2011	Sep 2011	Oct 2011	Nov 2011	Dec 2011	Jan 2012	Feb 2012
Sample 1	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8				
Sample 2				Q1	Q5	Q2	Q3	Q4	Q6	Q7	Q8	
Sample 3a					Q1	Q6	Q2	Q3	Q4	Q5	Q7	Q8
Sample 3b					Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8

Note: the questionnaires Q5 and Q6 do not follow the conventional order in samples 2 and 3a.

The assignment of incentives within the panel was varied experimentally during the recruitment interview; incentives were conditional on completing the wave questionnaire. For each questionnaire, respondents from sample 1 received either 5 or 10 Euros plus a 20-Euro bonus compensation for completing all eight questionnaires. Respondents from samples 2, 3a and 3b received 2 or 5 Euros, or no incentive, and no bonus payment for completing all questionnaires. Respondents could choose between having the money transferred to their bank account, sent to them as an online voucher, or having it donated to charity.

4.2.2 Survey evaluation measures and respondent characteristics

In order to measure survey experience during a particular wave, we asked evaluation questions at the end of every online questionnaire. We asked respondents to evaluate the survey as being “interesting”, “diverse”, “important for science”, “long”, “difficult”, and “too personal.” The wording of the items was developed with the help of the dictionary for survey evaluation based on the answers to the open question asking to rate the survey (Kaczmirek, Baier, & Zuell, 2010). Additionally, we included an item “important for science” to capture the importance of the survey sponsor. In previous research, academic/government sponsorship has been shown to enhance survey cooperation when compared to a commercial sponsor for cross-sectional surveys (Fox, Crask, & Kim, 1988; Heberlein & Baumgartner, 1978; Walston, Lissitz, & Rudner, 2006). For a panel study, Tortora (2009) also showed that “poll-like” surveys, which could be classified as social surveys, were less likely to suffer from attrition than studies classified as market research. The questions were asked as a single matrix with a fully labeled four-point scale (see Figure C1 in Appendix C).

Besides these specific aspects of the survey, we include in the analyses the overall evaluation of the questionnaire measured on a fully labeled five-point scale with a middle category. Another indicator that we use is the percent of item nonresponse within the questionnaire in the previous wave. Item nonresponse in the previous wave of the panel survey has been used as an indicator of a negative reaction of the respondent to the questions and/or generally negative atmosphere of the interview, and has been shown to increase unit nonresponse in the following survey (Loosveldt et al., 2002).

Incentive groups were comprised of those in the no-payment condition, groups of those receiving 2 Euros, 5 Euros, and two bonus conditions: 5 Euros plus bonus and 10 Euros plus bonus. The decision about the form of the incentive payment (bank transfer, coupon, transfer to charity) could be made as many times as a respondent decided to cash the incentive. Thus, various combinations of incentive payments are possible across respondents and on the respondent level. In order to avoid complications in the analyses, only a binary variable is included showing whether the respondent has ever initiated incentive payment (invariant of the form of payment).

We include the following respondent characteristics that have been shown to influence attrition in face-to-face studies: age, gender, educational level, employment status (working part or full-time coded as a reference category), immigration background, and living with a spouse or having a partner in a household. These characteristics were measured during the recruitment interview or during the second and the third wave for those in short interview conditions and are treated as time-invariant.

4.2.3 The definition of attrition and estimation models

The loss of panel members has multiple definitions. This is primarily due to the multitude of follow-up strategies used in panels to collect data from panel members who have been nonrespondents in one or multiple waves. When contrasted with a continuous participation of respondents in every wave, attrition can be defined as:

1. The first occurrence of wave nonresponse: all cases with at least one wave nonresponse are treated as attriters (e.g., Fitzgerald et al., 1998). Examples of the response pattern, where 0 stands for nonresponse in a particular wave and 1 for response, are: 11110000, 11101111, 10111111, etc.
2. Monotone attrition or occurrence of continued wave nonresponse up to the wave in question: only when a respondent never resumes participation is he or she considered to be an attriter (examples: 11111110, 11111000, 10000000, etc.).
3. Various patterns of wave nonresponse: occasional response, occasional nonresponse, and very irregular response (Nicoletti & Peracchi, 2005). Occasional response occurs when respondents take part in one or two subsequent waves (e.g., 11000000, 00010000, etc.), occasional nonresponse is the opposite case (e.g., 11110011, 10111111, etc.), very irregular response encompasses all other patterns.

Various patterns of wave nonresponse dictate the division of initial panel respondents into groups of attriters. For example, panelists who have been inactive for three months are called “sleepers” in the LISS Panel (Scherpenzeel & Zandvliet, 2011). Lugtig (2014) distinguishes nine classes of respondents according to their attrition behavior: loyal stayers, slow starters, gradual attriters who attrite after 40 months, after 30 months, after 18 months, or after 12 months, fast attriters, high lurkers, and lurkers (see also Lugtig, Das, & Scherpenzeel, 2014). In face-to-face household panels, monotone attrition prevails and irregular participation patterns constitute about ten percent of the cases (Nicoletti & Peracchi, 2005). Online panels typically involve frequent surveying and short questionnaires, making it easier to leave the panel or resume participation. They thus face greater challenges related to multiple response patterns (Lugtig, 2014). Within the GOPP, 255 different patterns (2^8-1) of nonresponse are theoretically possible for the first eight waves.

Strategies of modeling attrition are often based on the binary outcome variables for response or attrition. These strategies include: 1) response propensity models, usually logit or probit (e.g., Fitzgerald et al., 1998; Hawkes & Plewis, 2006; Laurie et al., 1999; Zabel, 1998); 2) dependent models of contact and cooperation with binary outcomes (Lepkowski & Couper, 2002); 3) sequential models, which examine response conditional on contact (Nicoletti & Peracchi, 2005; Watson & Wooden, 2009); 4) event history analysis with discrete time (Hawkes & Plewis, 2006; Lemay, 2009; Uhrig, 2008); 5) latent-class modeling (Lemay, 2009; Lugtig, 2014; Lugtig et al., 2014). Most of these models ignore the dynamic nature of attrition and the fact that attrition behavior of a respondent changes with the duration of the panel (De Keulenaer, 2005). We will employ a modeling strategy which accounts for time-varying covariates.

We use panel data analysis techniques to account for the longitudinal nature of survey experience. The dependent variable is a binary outcome, which takes the value of 0 for response in a particular wave and 1 in case of wave nonresponse. In this definition, we treat completion of a questionnaire as response, which means that persons who broke off during the questionnaire are treated as wave nonrespondents. We first fit a fixed effects logistic regression model (Allison, 2009, p. 28), which in its basic form can be expressed as:

$$log\left(\frac{p_{it}}{1-p_{it}}\right) = \mu_t + \beta x_{it} + \gamma z_i + \alpha_i + \varepsilon_{it}, \qquad t = 1, 2, \ldots, T.$$

In this model, p_{it}, is the probability to attrite, μ_t is an intercept which may be different for each period, x_{it} is a vector of time-varying predictors, z_i is a vector of time-invariant predictors, α_i and ε_{it} are two different error terms: α_i is an effect of all unobserved constant variables, whereas ε_{it} is different for each individual at each point in time (Allison, 2009, p. 6), and t is an indicator for the time period, that is, in our case for the panel waves. The advantage of the fixed effects model is that it implicitly controls for unobserved heterogeneity and only takes into account intrapersonal change.

The fixed effects regression model, however, cannot include predictors that are time-invariant, because fixed effects analysis only uses within-person variation. As our goal is to contrast survey experience (survey evaluation items measured in the previous wave and item nonresponse in the previous wave, which are time-varying) with incentives (assigned once, thus, time invariant), we also fit a random effects model, which can include both time-varying and time-invariant predictors:

$$log\left(\frac{p_{it}}{1-p_{it}}\right) = \mu_t + \beta x_{it} + \gamma z_i + \alpha_i + \varepsilon_{it}, \qquad t = 1, 2, \ldots, T,$$

where ε_{it} is random residual variation over time. The difference between fixed effects and random effects models is that in the random effects model α_i is assumed to have a specified probability distribution, whereas in the fixed effects model α_i it is treated as a set of fixed numbers, which is equivalent to treating it as random with all possible correlations with x_{it} (Allison, 2009, p. 21). The assumption that is uncorrelated with all other variables is strong and may lead to the model producing estimates that are more biased than the estimates in the fixed effects model. To take into account the differences in attrition over the course of the panel, such as that attrition is highest after the first wave (Fitzgerald et al., 1998) or the first few waves (Uhrig, 2008) and then flattens out, we include period (wave) indicators into the regression models. Since we use panel data for our analysis, that is, multiple cases are produced by the same individual, observations are not independent. This means that the errors are potentially correlated. Therefore, we report clustered standard errors that account for the correlation of observations within one individual.

Both fixed effects and random effects models have the drawback that they only estimate transitions from response to nonresponse. That is, we do not include the transition from wave nonresponse to wave nonresponse in the analysis. One possible solution would be to impute the values of survey evaluation and item nonresponse percentage for the waves that respondents skipped. However, imputation presupposes that we rely on the model which relates unobserved values to observed values and the reasons for missing data are correctly modeled (Rubin, 1978, 1996). Modeling of the panel nonresponse process, that is, the process which produces missing estimates, is the objective of this chapter. Therefore, it would make little sense to impute the values for survey evaluation and item-nonresponse in this particular case.

We account for multiple wave nonresponse by fitting a multinomial logistic regression model for the following groups of attriters: monotone attrition (patterns 11111110, 11111100, …, 11000000), irregular participation (all other patterns excluding the above and 10000000), and one-time only participants (pattern 10000000). The reference category is the continued participation with a pattern (11111111). This classification resembles the theoretical classification of Lemay (2009)[16], who differentiates between one time participants/early attriters, repeated participants/late attriters, and stayers (p. 102), and an empirical classification of Lugtig (2012) with four classes: fast attriters, lurkers, gradual attriters, and loyal stayers. Multinomial logit does not account for timing of attrition and discards the longitudinal character of the data, however, it allows to differentiate between attrition patterns.

4.3 Results

Among the 1643 individuals in our sample, 37.25% never participated in the panel and 33.48% completed all waves. The rest of the panelists completed one (9.25%) or several waves of the panel (20.2%). Of those who have participated in at least one wave, 46.65% became wave nonrespondents at some point. The maximum attrition rate as defined by DiSogra and Callegaro (2009) is:

$$Maximum\ attrition\ rate = \frac{first\ wave\ respondents - all\ waves\ respondents}{first\ wave\ respondents} = \frac{1000-550}{1000} = 45\%.$$

The first figure (46.65%) is somewhat higher because the panel did not have a profile survey and respondents did not have to complete the first wave in order to continue to participate

16 Due to the computational difficulties and model complexity, he could not empirically verify his proposed classification.

in the panel. A small number of participants never completed the first online wave of the panel. In Table 4.2, we report the response and attrition statistics across eight waves of the GOPP.[17] We use the following definition of attrition:

$$\textit{Wave-based attrition} = \frac{\textit{Ever active respondents} - \textit{respondents wave}_t}{\textit{Ever active respondents}}.$$

This comes close to the baseline-on-wave attrition rate (DiSogra & Callegaro, 2009):

$$\textit{Baseline-on-wave attrition rate} = \frac{\textit{Respondents@baseline} - \textit{respondents@wave}_t}{\textit{Respondents@baseline}}.$$

However, the former formula allows to account for the design specifics of GOPP that did not require the panelists to start with the first online wave. The percent of panelists invited to waves one through eight decreases as a result of panel members signing off from the panel. We do not differentiate here between the panelists actively signing off from the panel and those staying with the panel but not responding to a particular wave. All respondents who completed at least one questionnaire within the panel are treated as eligible for the following waves and coded as 1 on wave nonresponse even when they actively quit participating. For the calculation of attrition, response to wave *t* is defined as the number of completes for this wave.

Table 4.3 presents the between-person and within-person variances for the survey evaluation items and the dependent variable of wave nonresponse. For survey evaluation, the within-person variance is greater than the between-person variance suggesting that fixed effects regression is the right course of action for studying the effects of survey experience, because the fixed effects model is better suited for taking into account the within-person variation. The between-person variation is greater for wave nonresponse than the within-person variation, which is not surprising given that most of the cases were complete respondents.

Table 4.2. Response and attrition statistics by wave.

	Wave 1	**Wave 2**	**Wave 3**	**Wave 4**	**Wave 5**	**Wave 6**	**Wave 7**	**Wave 8**
Overall N invited	1643	1636	1618	1593	1578	1569	1560	1556
Percent started if invited	60.86	51.89	50.19	49.34	47.53	46.34	46.54	45.76
N started	1000	849	812	786	750	727	726	712
N completed	925	804	782	767	742	700	715	682
Wave-based attrition, %	10.28	22.02	24.15	25.61	28.03	32.10	30.65	33.85

Note: N active panelists over the course of the entire panel is 1031.

[17] Note that attrition metrics reported here are based on the analysis sample and would be different if calculated with nonresponse follow-up respondents included.

Table 4.3. Means and standard deviations for survey evaluation items and wave nonresponse.

Variable	Mean	SD Overall	SD Between	SD Within	N observations	N persons
Interesting	3.006	0.714	0.484	0.558	6087	1025
Diverse	2.862	0.713	0.462	0.575	6090	1025
Important for science	2.811	0.720	0.522	0.539	6049	1024
Long	1.847	0.714	0.502	0.562	6081	1025
Difficult	1.652	0.708	0.454	0.570	6080	1021
Too personal	2.052	0.838	0.550	0.665	6090	1024
Overall evaluation	3.678	0.734	0.498	0.570	6099	1030
Wave nonresponse	0.535	0.499	0.449	0.217	13144	1643

Note: items "interesting," "diverse," "important for science," "long," "difficult," "too personal," and "overall evaluation" were measured directly, the indicator "wave nonresponse" is constructed.

4.3.1 Characteristics of those who attrite

Table 4.4 presents descriptive statistics and tests of significance for the stable and time-varying respondent characteristics wave by wave. All variables treated as time-invariant are compared for each wave to wave 8. Survey experience variables, including item nonresponse are compared for each wave to the previous wave. Generally, the sample composition does not change much in the course of the panel. In none of the waves are demographic characteristics significantly different from the final wave. The overall change in the composition of the incentive groups seems to be small. However, if we calculate retention rates as the share of those who completed the eighth wave over those who completed the first wave for each group, the results are: 65% for the no-incentive group, 69.2% for the 2 Euros incentive group, 71.93% for the 5 Euro-group, 81.37% for the 5 Euro-group with a bonus, and 84.21% for the 10-Euro group with a bonus.

The survey evaluation items and item nonresponse vary significantly from wave to wave but there is no clear trend. One example of a clear trend indicative of intrinsic motivation would be the increase of the overall evaluation and increase in rating the survey as interesting. An alternative trend would be a decrease in means of the items "long," "difficult," and "too personal" if those who find the survey too burdensome stop participating in the panel. Multivariate analysis will allow us to study the relationship between evaluation of the previous wave and incentives with panel attrition.

Table 4.4. Respondent characteristics, incentive groups, and survey evaluation by wave.

Variable	Wave 1	Wave 2	Wave 3	Wave 4	Wave 5	Wave 6	Wave 7	Wave 8
Age	42.689 (14.561)	43.177 (14.622)	43.065 (14.562)	43.423 (14.677)	43.465 (14.710)	43.415 (14.914)	43.562 (14.719)	43.521 (14.741)
Male	53.51	53.11	51.41	52.41	50.94	51.86	52.17	52.47
Education: low	11.12	11.04	9.95	10.28	9.25	9.52	9.20	9.45
Education: middle	28.33	27.86	28.53	26.70	28.31	27.67	27.30	27.89
Education: high	60.55	61.10	61.52	63.02	62.43	62.81	63.51	62.67
In paid work	71.82	73.19	72.08	71.31	71.47	71.03	71.71	71.83
Immigration background	9.77	10.52	9.42	9.52	9.37	9.20	9.56	9.62
Married/has partner	69.83	68.98	69.01	68.95	68.54	68.66	68.98	69.08
Incentive: 0	17.30+	16.79	15.86	16.43	15.63	15.71	15.10	15.25
Incentive: 2	24.22+	23.38	23.53	22.69	23.72	22.71	23.22	22.73
Incentive: 5	24.65+	24.00	24.55	24.64	23.85	24.29	24.06	24.05
Incentive: 5+20	17.41+	18.03	18.41	18.38	19.00	18.71	19.02	19.21
Incentive: 10+20	16.43+	17.79	17.65	17.86	17.79	18.57	18.60	18.77
Overall evaluation	3.746 (0.597)	3.616*** (0.698)	3.732*** (0.663)	3.764 (0.717)	3.747 (0.711)	3.402*** (0.897)	3.628*** (0.807)	3.758** (0.724)
Interesting	2.978 (0.625)	2.948 (0.706)	3.015+ (0.685)	3.037 (0.714)	3.067 (0.704)	2.805*** (0.799)	3.014*** (0.754)	3.196*** (0.683)
Diverse	2.996 (0.565)	2.745*** (0.724)	2.798 (0.687)	2.890** (0.681)	2.923 (0.671)	2.555*** (0.840)	2.790*** (0.709)	3.193*** (0.662)
Important for science	2.815 (0.660)	2.782 (0.693)	2.876** (0.665)	2.876 (0.727)	2.912 (0.714)	2.547*** (0.792)	2.847*** (0.754)	2.820 (0.707)
Long	1.829 (0.660)	1.864 (0.712)	1.742*** (0.635)	1.673* (0.593)	1.745* (0.671)	2.000*** (0.814)	1.724*** (0.602)	2.250*** (0.850)
Difficult	1.408 (0.545)	1.503*** (0.599)	1.486 (0.599)	1.563* (0.657)	1.759*** (0.741)	1.597*** (0.662)	1.948*** (0.812)	2.073** (0.777)
Too personal	1.695 (0.634)	2.205*** (0.811)	2.160 (0.823)	1.919*** (0.768)	1.804** (0.718)	2.451*** (0.940)	2.062*** (0.896)	2.222*** (0.861)
INR in percent	0.398	0.676***	0.438**	0.159***	0.153	0.199	0.326	0.225

Note: means and percentages are reported, standard deviations in parentheses, INR is short for item nonresponse. Due to the changed order of administering questionnaires, survey evaluation items reflect content evaluation of questionnaire content for waves 1, 7, and 8. For group sizes see more elaborate Table C1 in Appendix C. +p<0.1, *p<0.05, **p<0.01, ***p<0.001.

4.3.2 Explaining attrition: The role of survey experience and incentives

Table 4.5 summarizes the results of two estimation models: the fixed effects model and the random effects model, both using a binary dependent variable “nonresponse in the following wave.”

Table 4.5. Fixed effects and random effects models of panel attrition.

	Fixed effects model			Random effects model		
Predictor variables	**Odds Ratio**	**Robust SE**	**p-level**	**Odds Ratio**	**Robust SE**	**p-level**
t-1: interesting	1.006	0.160	0.969	1.163	0.175	0.316
t-1: diverse	1.093	0.171	0.571	1.003	0.155	0.987
t-1: important for science	0.771	0.107	0.059	0.713	0.090	0.007
t-1: long	1.046	0.140	0.734	1.178	0.159	0.224
t-1: difficult	1.008	0.141	0.953	1.069	0.138	0.605
t-1: too personal	1.277	0.142	0.028	1.250	0.132	0.034
t-1: overall	1.070	0.163	0.657	0.871	0.128	0.347
t-1: INR	0.049	0.312	0.637	24.335	136.047	0.568
Age				0.986	0.009	0.132
Male				1.332	0.343	0.265
Education: low				reference		
Education: middle				0.469	0.203	0.081
Education: high				0.249	0.104	0.001
In paid work				1.095	0.313	0.751
Immigration background				1.532	0.650	0.315
Married/has partner				1.210	0.329	0.484
Incentive: 0				reference		
Incentive: 2				0.297	0.112	0.001
Incentive:5				0.201	0.077	0.000
Incentive: 5+20				0.083	0.036	0.000
Incentive: 10+20				0.068	0.030	0.000
Wave 2	reference			reference		
Wave 3	1.665	0.397	0.032	0.566	0.146	0.027
Wave 4	1.773	0.531	0.056	0.585	0.150	0.037
Wave 5	2.843	0.880	0.001	0.774	0.204	0.331
Wave 6	5.887	1.956	0.000	1.096	0.316	0.751
Wave 7	3.092	1.169	0.003	0.532	0.171	0.049
Wave 8	5.644	2.140	0.000	0.913	0.263	0.751
Constant				0.988	0.935	0.990
N observations	1128			4388		
N persons	250			810		
Wald χ^2	55.69		0.000	86.84		0.000
Log pseudolikelihood	-347.566			-1144.096		
Rho				0.691	0.046	
lnsigu^2				1.997	0.217	
sigma_u				2.715	0.298	

Note: The number of observations/persons in the fixed effects model is restricted to cases that experience a transition from response to nonresponse, other cases are discarded according to the estimation logic. Pseudo-R^2 for the fixed effects model is 0.10. Rho represents the estimated residual intraclass correlation, sigma_u is the estimated residual standard deviation of the random intercept. SE is short for standard errors.

Among the survey experience items, two items predict attrition. Rating the previous survey as important for science decreases the chances of nonresponse in the current wave (Odds Ratio 0.771, $p<0.10$), rating the previous survey as too personal increases the chances of wave nonresponse by roughly one third (Odds Ratio=1.277, $p=0.028$). In the random effects model, which in addition to survey experience indicators includes time-invariant covariates, both survey evaluation measures "important for science" and "too personal" demonstrate a pattern similar to the fixed effects model. Among the demographic variables, only education is significantly related to attrition: those at the highest educational level are less likely to attrite. There are significant effects for the amount of incentives with conditional odds decreasing as the incentive amount increases. The most substantial difference in the incentive groups occurs between the conditions with bonus and those without bonus: conditional odds[18] of attrition decrease by 70.3% and by 79.9% for non-bonus groups and by 91.7% and 93.2% in the bonus conditions.

Despite the capability of the random effects model to include time-invariant covariates, the estimates produced by the random effects model can be biased when compared to fixed effects estimates (Allison, 2009). Particularly problematic in our case is the estimate of item nonresponse, which has unusually large standard errors. Therefore, we also used the hybrid model. The hybrid model combines the advantages of the fixed effects and random effects models (Allison, 2009, p. 23). The hybrid model allows for obtaining estimates for the time-varying covariates calculated in the same manner as in fixed effects models, and also allows for the inclusion of time-invariant covariates as random effects models do. The three types of models (fixed effects, random effects, and hybrid) allow us to focus on different aspects of our research questions. The fixed effects model is best suited for studying the influence of the time-varying survey experience measures on attrition. The random effects model can estimate the effects of incentives and other time-invariant predictors, such as gender and education. The hybrid model provides less biased estimates for the time-varying predictors than the random effects model would, while allowing for the inclusion of time-invariant predictors, which is impossible in the fixed effects model.

The results of the hybrid model are shown in Table 4.6 (Model 1). As in the fixed effects model, all time-varying variables are expressed as deviations from their person-specific means. In Table 4.6 the coefficients for the deviations ("D") and the means ("M") are reported. The coefficients for the deviations are interpreted later in this section. The means are less interesting to interpret, but they need to be included in the model.

In the hybrid model the effects found earlier hold for the items "important for science," which is negatively related to attrition, and "too personal," which predicts attrition. Additionally, rating the previous survey as difficult increases the likelihood of attrition in the current wave by 48%.

Since in both the fixed effects and the random effects models the Odds Ratio for "difficult" was not much higher than one, meaning that the likelihood of attrition is roughly the same for respondents with different levels of rating the survey as difficult, this is an unexpected result. It will be further discussed in the analysis of participation patterns (see section 4.3.4).

[18] Calculated as 100%(Odds Ratio–1).

Table 4.6. Hybrid models of panel attrition.

	Model 1			Model 2 (with payment)		
Predictor variables	**OR**	**Robust SE**	**p-level**	**OR**	**Robust SE**	**p-level**
D: interesting	1.171	0.184	0.315	1.135	0.212	0.495
D: diverse	0.984	0.164	0.925	0.951	0.194	0.807
D: important for science	0.746	0.104	0.035	0.754	0.123	0.083
D: long	0.976	0.142	0.866	0.876	0.151	0.443
D: difficult	1.481	0.207	0.005	1.463	0.260	0.032
D: too personal	1.499	0.167	0.000	1.622	0.231	0.001
D: overall	0.974	0.155	0.867	1.088	0.222	0.678
D: INR	5.685	32.339	0.760			
M: interesting	1.676	0.804	0.282	1.125	0.454	0.770
M: diverse	1.056	0.464	0.902	0.633	0.253	0.252
M: important for science	0.523	0.171	0.048	0.781	0.230	0.401
M: long	4.311	1.535	0.000	2.417	0.789	0.007
M: difficult	0.193	0.071	0.000	0.474	0.157	0.024
M: too personal	0.432	0.129	0.005	0.856	0.232	0.566
M: overall	0.480	0.214	0.099	1.150	0.452	0.722
M: INR	0.000	0.000	0.042			
Age	0.981	0.009	0.043	0.964	0.010	0.000
Male	1.289	0.327	0.316	1.003	0.254	0.993
Education: low	reference			reference		
Education: middle	0.512	0.217	0.114	0.736	0.295	0.444
Education: high	0.293	0.119	0.002	0.642	0.244	0.244
In paid work	1.091	0.308	0.759	1.063	0.304	0.832
Immigration background	1.573	0.648	0.271	1.021	0.405	0.958
Married/has partner	1.092	0.289	0.740	1.111	0.299	0.695
Incentive: 0	reference					
Incentive: 2	0.392	0.142	0.010	reference		
Incentive:5	0.256	0.094	0.000	1.223	0.398	0.536
Incentive: 5+20	0.115	0.048	0.000	0.663	0.232	0.240
Incentive: 10+20	0.103	0.044	0.000	0.896	0.329	0.766
Payment				0.011	0.005	0.000
Wave 2	reference			reference		
Wave 3	0.503	0.130	0.008	0.521	0.159	0.032
Wave 4	0.539	0.138	0.016	0.616	0.178	0.093
Wave 5	0.719	0.189	0.210	0.798	0.254	0.479
Wave 6	0.995	0.287	0.985	1.291	0.436	0.449
Wave 7	0.493	0.158	0.028	0.545	0.209	0.114
Wave 8	0.746	0.220	0.319	0.897	0.305	0.749

Table 4.6. Continued.

	Model 1			Model 2 (with payment)		
Predictor variables	**OR**	**Robust SE**	**p-level**	**OR**	**Robust SE**	**p-level**
Constant	75.071	137.752	0.019	7.554	12.969	0.239
N observations	4388			3764		
N persons	810			669		
Log pseudolikelihood	-1111.226			-714.532		
lnsig2u	1.831	0.205		1.185	0.271	
sigma_u	2.499	0.256		1.808	0.245	
Rho	0.655	0.046		0.498	0.068	
Wald χ^2	124.28		0.000	176.50		0.000

Note: when tested against random effects model, 4 of the time-varying covariates show significant differences and 4 coefficients do not, overall Wald test is against random effects model: $\chi^2(8)= 46.49$, $p< 0.001$ and in favor of the hybrid model. "D" means coefficients for the deviation, "M" means coefficients for the means.

Among demographic characteristics, the highest educational level is negatively related to attrition and the effect is statistically significant. In the hybrid model the effect of age becomes statistically significant, but nevertheless remains rather small: the conditional odds of attrition decrease by 1.9% with every additional year of respondent's age. For the incentive groups the results hold: the higher the incentive, the lower the probability of attrition in the following wave. There is a significant difference when we compare groups that are paid to those who were offered no incentive. Among differential incentive amounts there is no difference between the 2-Euro group and the 5-Euro group. If the incentive offer of 2 Euros is treated as a reference, the incentives conditions of 0 Euros, 5 Euros plus bonus, and 10 Euros plus bonus remain significant (OR=2.552, p=0.010; OR=0.294, p=0.002; OR=0.264, p=0.001 respectively and OR=0.653, p=0.225 for 5-Euro group without bonus, full results not shown).

Overall, we find some support for Hypothesis 1, which predicted that positive panel experience inhibits attrition. There is evidence that a *burdensome experience increases the likelihood of further nonparticipation*, however, it is not confirmed that positive experience with the panel inhibits attrition. Furthermore, *perceiving previous surveys as important for science decreases the likelihood of nonparticipation* in the following waves. We find support for Hypothesis 2, which expected incentives to decrease the likelihood of attrition. The analyses show that *being in the incentive group as opposed to receiving no incentives decreases the likelihood of attrition*. The analyses also show that *the higher the offered incentive, the less likely the respondent is to attrite*. However, incentives offered in GOPP were not automatically paid to the respondent but needed to be redeemed and the respondent could choose the form – cash, voucher, or donation to charity. Therefore, to test Hypotheses 3 and 4 – whether incentives are an additional bonus to the positive survey experience or act as compensations for negative survey experiences – we need to include information about incentive redemption into the model.

4.3.3 The interplay between survey experience and incentives

Model 2 in Table 4.6 includes an indicator variable whether a panel respondent has initiated incentive redemption in the course of the panel or not. Here, the form of the incentive (cash, voucher, or donation) is not included, because respondents could switch between the forms any time during the waves and did that in an unsystematic way. Respondents did not have to redeem their incentive after each wave, but could choose to cash it any time they wanted: after each wave, after completing several waves, at the end of the eight panel waves, etc. It makes little sense to arbitrary assign the redemption to any particular wave and treat this indicator as a time-varying variable, because respondents could freely choose when and how to cash in their incentive. For example, one respondent initiated redemption 27 times over the course of the panel in a relatively short time span. If we recoded his or her redemption behavior as time-varying indicators for each wave, some waves would be coded "0" while other waves in which the incentives for the previous waves were also redeemed would be coded "1". These indicators would not reflect the payment for the respective waves. Therefore, the indicator of payment is binary (1 for "ever redeemed incentive" and 0 for "never initiated redemption). After we control for incentive redemption, the effects of the amounts of incentive are no longer statistically significant. For respondent attrition it does not make a difference to which incentive group the respondent belonged, as long as he or she cashed in the incentive. It may however be the case that those offered larger incentives are more likely to go through the trouble of collecting the incentive payment. To test for this, we analyzed the interactions of incentive amount with redemption and also looked into which respondents are more likely to redeem their incentives. Overall, the likelihood of incentive redemption increases with the incentive amount: Odds Ratio=2.488 for a 5-Euro group, 2.299 for a 5-Euro plus bonus group and 4.083 for a 10-Euro plus bonus group (all $p<0.01$) and almost three times with every completed survey (Table C2 in Appendix C). None of the interactions of incentive amounts with redemption are significant (Table C3 in Appendix C). Therefore, *although generally incentive amounts influence the likelihood of incentive redemption, they do not influence attrition once incentives are redeemed*. Likewise, neither of the interactions of incentive redemption with survey experience is statistically significant (Table C4 in Appendix C).

Thus, incentives do not compensate for negative experience within the panel, which was expected by Hypothesis 3. The Hypothesis 4, according to which the incentives are not important when the experience within the panel is positive, can be discarded as well. *Incentives seem to be an additional bonus to the overall positive survey experience and for those rating the surveys as important for science.*

4.3.4 Participation patterns

Overall, 1031 respondents show 108 different response patterns. Based on the review of the various attrition definitions in section 4.2.3, we define four participation patterns. For each pattern, "1" represents response and "0" represents wave nonresponse. For a three-wave panel, a pattern "101" means that the respondent completed the first wave, did not respond to the second wave and responded to the third wave. We analyze the panel with eight waves, so each pattern has eight digits. We distinguish between four participation patterns, three of which are attrition patterns: 1) continued participation (pattern 11111111); 2) monotone attrition (patterns 11111110, 11111100, …, 11000000); 3) irregular participation (all other patterns excluding the above and 10000000) and 4) first wave only participants (pattern 10000000). Continued participation is the most common pattern: 53.93% of panel members take part in all of the surveys. The next most common pattern is irregular participation with 24.35% of participants. Monotone attrition and one time only participants that also can be seen as a subgroup of monotone attrition each amount to 10.86% of participants.

A multinomial logistic regression was performed with the dependent variable "attrition group" and the reference group of continued panel participation. The results are summarized in Table 4.7. According to Hypothsis 5, respondents motivated by positive experience and respondents motivated by incentive payment should have dissimilar participation patterns. The relative risk ratios (RRR) are obtained by exponentiating the multinomial logit coefficients and can be interpreted as Odds Ratios (Rabe-Hesketh & Skrondal, 2012, p. 637). The values of relative risk ratios for survey experience in Table 4.7 show that two groups of attriters are similar: first wave only participants and monotone attritters. Respondents who rate the survey as more diverse, less important for science, long, more difficult and too personal are more likely to be monotone attriters or first wave only participants than to be committed respondents who take part in every wave. Respondents who participate irregularly are more likely to rate the surveys higher on the diversity, but their ratings of the surveys as important for science and burdensome do not differ from the group of respondents who participate in every wave. In terms of survey experience, respondents with irregular participation resemble respondents who always participate, whereas those who attrite monotonically resemble those who only take part in the first wave.

It is expected that across all attrition groups the relative risk ratios decrease as the incentive amount increases. This is indeed the case with the exception of the monotone attrition group receiving 10 Euros plus bonus, whose relative risk ratio is higher than the 5-Euro plus bonus condition. Overall, those offered incentives are less likely to be in an attrition group than in a group of continued participation. The effects are statistically significant except for the 2-Euro monotone attrition group. Furthermore, the higher the offered incentive, the lower is the relative risk ratio to be in an attrition group relative to the continued participation group.

Among the demographic variables, we see a lower relative risk ratio for older first-wave-only participants (RRR=0.975) and a substantively higher risk ratio for those in paid work to be in a monotone attrition group than in continued participation group (RRR=2.172). Education, which was negatively related to attrition on a statistically significant level, does not predict attrition group membership relative to continued participation. Collectively, respondents who are demotivated by a negative survey experience differ in their participation patterns from those who participate in every wave. *Respondents who rate previous waves as long, difficult, or too personal have a higher chance to attrite monotonically after the first or after several panel waves.* Participants with irregular participation patterns do not differ from the group of continued

Table 4.7. Multinomial logistic regression with continued participation as a reference category. RRR indicates the Relative Risk Ratio

	Only first wave			Irregular participation			Monotone attrition		
Predictor variables	**RRR**	**SE**	**p-level**	**RRR**	**SE**	**p-level**	**RRR**	**SE**	**p-level**
D: interesting	1.017	0.026	0.512	1.002	0.039	0.967	1.051	0.029	0.078
D: diverse	1.193	0.033	0.000	1.129	0.039	0.000	1.204	0.032	0.000
D: important for science	0.937	0.018	0.001	0.976	0.032	0.457	0.926	0.018	0.000
D: long	1.203	0.030	0.000	1.041	0.040	0.293	1.181	0.035	0.000
D: difficult	1.222	0.036	0.000	1.060	0.036	0.088	1.184	0.027	0.000
D: too personal	1.047	0.018	0.007	1.026	0.025	0.304	1.067	0.022	0.002
D: overall	1.042	0.024	0.076	0.992	0.035	0.835	1.013	0.023	0.562
M: interesting	0.475	0.360	0.325	1.232	0.464	0.580	0.715	0.369	0.516
M: diverse	2.243	1.546	0.241	0.724	0.260	0.368	1.074	0.442	0.861
M: important for science	1.345	0.569	0.484	0.820	0.221	0.461	0.754	0.252	0.397
M: long	3.920	1.775	0.003	1.946	0.557	0.020	1.789	0.679	0.125
M: difficult	0.085	0.054	0.000	0.818	0.233	0.481	0.344	0.146	0.012
M: too personal	0.426	0.184	0.048	0.780	0.201	0.336	0.846	0.243	0.561
M: overall	0.471	0.280	0.205	0.657	0.234	0.238	0.805	0.401	0.663
Age	0.975	0.010	0.013	1.002	0.007	0.822	0.988	0.010	0.207
Male	1.722	0.562	0.096	0.907	0.185	0.630	0.742	0.194	0.253
Education: low	Ref.			Ref.			Ref.		
Education: middle	0.686	0.336	0.441	1.019	0.368	0.959	0.567	0.250	0.197
Education: high	0.532	0.233	0.150	0.729	0.254	0.364	0.531	0.206	0.102
In paid work	1.322	0.494	0.455	1.258	0.283	0.309	2.172	0.700	0.016
Immigration background	0.812	0.478	0.723	1.291	0.448	0.461	1.254	0.508	0.576
Married/has partner	1.114	0.378	0.750	0.871	0.199	0.544	0.759	0.225	0.352
Incentive: 0	Ref.			Ref.			Ref.		
Incentive: 2	0.358	0.152	0.015	0.553	0.165	0.046	0.644	0.237	0.233
Incentive:5	0.268	0.115	0.002	0.352	0.109	0.001	0.381	0.153	0.016
Incentive: 5+20	0.197	0.091	0.000	0.270	0.091	0.000	0.296	0.133	0.007
Incentive: 10+20	0.154	0.076	0.000	0.249	0.086	0.000	0.341	0.149	0.014
Constant	19.600	40.644	0.151	4.022	6.189	0.366	15.519	33.292	0.201

Note: N observations=4039, N persons=809, Log pseudolikelihood -2942.654, Wald χ^2=339.46, p<0.001, pseudo-R^2=0.075. Coefficients test: χ^2 (24)=53.43, p<0.001; Standard errors (SE) are robust standard errors. The variable percent of item nonresponse in the previous wave is excluded due to the small number of observations in the group of first wave only participation. "D" means coefficients for the deviation, "M" means coefficients for the means.

participation on the evaluation of the previous waves with the exception of the item "diverse." Offering incentives, on the other hand, decreases the likelihood to be in an attrition group rather than in an "always in" continued participation group for all incentive levels.

This and the other analyses presented above suggest that respondents are motivated both intrinsically by non-burdensome survey experience, as indicated by their evaluations of the previous panel wave, and extrinsically by offered (monetary) incentives. As it was mentioned earlier in this chapter, we asked panel respondents whether they would like to continue to participate after the eighth wave. The results of a logistic regression comparing those who would like to continue and those who would like to quit the panel show that those who overall rated the surveys more positively and those who were offered incentive payments wanted to continue panel participation (Table C5 in Appendix C). This provides further support for the duality of the motivation not to attrite.

4.4 Conclusions and discussion

In this chapter, we apply the benefit-cost theory of survey participation (Singer, 2011) to study panel attrition. Survey experience in the previous panel wave may be either attributed to the costs of participation when the participation is burdensome, or to the benefits when the participation experience is a pleasant one. We contrast intrinsic motivation of a pleasant survey experience in the previous wave to the extrinsic motivation of incentives. We test whether incentives are an additional benefit to the pleasant survey experience, a compensation for an unpleasant survey experience, or an independent motivation factor. The data come from eight waves of a probability-based online panel of Internet users in Germany, in which differential incentives were assigned experimentally during the recruitment interview. We concentrate on the cases who participated in at least one wave of the panel to study how survey experience, offered incentives, and respondent characteristics shape panel attrition. The content of the surveys differed between the waves providing more variation in respondent evaluations of the surveys.

Our primary focus is on the time-varying survey experience variables which were measured in every wave of the survey. This distinguishes our analysis from the common approach of studying panel attrition, when the characteristics used as predictors of panel attrition are measured during the first panel wave. In contrast, we look at every transition from response to nonresponse. We find that survey experience of the previous wave influences the subsequent response behavior. Respondents who rate the surveys as important for science are less likely to attrite. Respondents for whom the experience is burdensome – too personal, too difficult – are more likely to become nonrespondents in the next survey wave. Offering monetary incentives inhibits panel attrition. However, the incentive amounts do not influence the attrition behavior after statistically controlling for the incentive redemption. Nevertheless, those respondents who are assigned higher amounts and who complete more surveys are more likely to redeem their incentives. There is no indication that incentives serve as a compensation for a burdensome survey experience.

Due to the nature of our research question, we cannot include in the analysis multiple nonresponse waves as respondents could not evaluate a (previous) survey in which they have not participated. To account for the presence of continued occurrence of wave nonresponse, we analyze the attrition determinants in three attrition groups which we constructed based on the participation patterns and the literature. We contrast the groups of fist-wave-only participants, respondents with irregular participation patterns, and respondents who attrite monotonically to the group of respondents with continued participation. We find that the groups of monotone attriters and first-wave-only respondents are similar in that their attrition behavior is influenced by the response burden. That is, the evaluations of the surveys as too personal, difficult and long increase the likelihood of respondents having those attrition patterns opposed to having a continued participation pattern. The same holds for evaluating the previous survey as too "diverse." Evaluating the surveys as important for science decreases the likelihood of panel members becoming nonrespondents after one or several waves of the panel. For the group of irregular participants, survey evaluation measures do not differ from those who take part in every panel wave. Incentives decrease the likelihood of being in either attrition group, the higher the amount, the lower the chances.

We do not find that demographic characteristics of respondents play such an important role as has been shown in other panel attrition studies. This may have several reasons: the panel from which we analyze the data being an online-administered panel survey (whereas the majority of studies on panel attrition investigate attrition in interviewer-administered panel surveys), the short intervals between the waves, or the fact that demographic characteristics may only serve as proxies of the actual determinants of panel attrition. Our results are in line with the findings on demographics being generally not predictive of attrition in other probability-based online panels. For the Knowledge Panel with weekly questionnaires, Clinton (2001) found older panelists to be more likely to attrite. Other demographic characteristics were not statistically significant. For the LISS Panel with monthly questionnaires, Lugtig (2014) reports that older respondents and lower educated respondents are more likely to attrite, whereas gender, urbanicity, or having a partner are not predictive of attrition. The reasons for demographics not playing a major role have yet to be investigated. In any case, this finding has important implications for constructing attrition weights for online panels, because attrition weights are usually based on demographic characteristics. The evidence presented here on the importance of survey design features (incentives) and evaluation of the survey experience suggests that these variables need to be included into the calculation of attrition weights.

Our analyses have several limitations which should be addressed in future studies. Given the relatively small number of cases (for panel analysis methods such as fixed effects models or the multinomial regression model), we do not differentiate between the panelists who actively sign off from the panel and those who continue to be invited but do not participate. In some studies it is argued that it is important to distinguish these two types of refusals: wave-nonresponse and withdrawal from the survey (Hawkes & Plewis, 2006; Laurie et al., 1999). Therefore, our analyses should ideally be replicated on a larger dataset with including withdrawal from the panel either as a control variable or as a separate attrition group. Furthermore, we concentrate on the refusal part of modeling attrition, assuming that respondents who provided valid email addresses at the recruitment stage can be contacted. Future studies would benefit looking in detail also on noncontact in online panel surveys.

Lastly, we do not differentiate here between various forms of incentive redemption. Respondents may differ in their preferences of incentive forms. For example, elsewhere we have shown that men and women prefer different incentive forms (Schaurer, Struminskaya, & Kaczmirek, 2014). Respondents choosing the form of donation to charity might differ from those choosing cash or coupons also on the dimensions of survey experience. The analysis of the choice of the incentive form (material form of cash or voucher vs. donation) can provide further insight into the respondent's motivation. Furthermore, it would be beneficial to include time-varying demographic covariates (such as changes in employment status or family situation) if these are available. We were unable to do so because of the design decision in the GESIS Online Panel Pilot to collect the primary demographic information during the recruitment interview and supplement it with the information provided in the later waves. The respondents were not asked to update their demographic information every month since it would have increased the respondent burden and the surveying time. In online panels, which allow respondents to update their demographic information or which explicitly ask to report changes regularly, including these variables is possible. However, one needs to be aware of the possible selection effects. It might be the case that not every respondent would report changes. Moreover, if attrition is a result of such changes, demographic updates would not capture that, only the analysis of validation data can lead to such conclusions.

Overall, as we have shown for survey experience measures, inclusion of dynamic factors measured directly and indirectly (such as item nonresponse in our case) benefits the analysis of panel attrition. One practical implication for survey researchers collecting panel data would be to measure those indicators in every wave.

Chapter 5

Respondent conditioning in online panel surveys: Results of two field experiments

A shorter version of this chapter has been conditionally accepted for publication as Struminskaya, B. Respondent conditioning in panel surveys: Results of two field experiments. *Social Science Computer Review – Special Issue on Survey Error in Probability-based Online and Mixed-Mode Access Panels of the General Population*

Abstract

In this chapter, changes in reporting due to prior interviewing are investigated. Two field experiments were implemented in a probability-based online panel in which the order of the questionnaires was switched. Although experimental methods for studying panel conditioning are favorable, experiments in longitudinal studies are rare. Studies on conditioning demand additional resources and might influence respondents' answers, which would be problematic. Panel conditioning is mostly associated with measurement errors. However, the discussion that sees it exclusively as a negative phenomenon is not comprehensive. Learning the rules that govern the interview may lead to increases or decreases in data quality (advantageous vs. disadvantageous conditioning). Overall, little evidence of advantageous conditioning and no disadvantageous conditioning is found. Apart from this reassuring finding for data users, this chapter advances the field by using propensity score weighting to account for attrition and other confounding factors and by relying on paradata to rule out alternative explanations of panel conditioning.

5.1 Introduction

Panel conditioning is defined as a learning effect that occurs when participation in previous survey rounds changes attitudes and behavior or the way they are reported in subsequent interviews (Kalton, Kasprzyk, & McMillen, 1989; Waterton & Livesley, 1989). In terms of data quality, conditioning can be advantageous ("good") or disadvantageous ("bad"). Advantageous conditioning denotes increased quality of reporting through better understanding of the survey procedure or survey instrument. Disadvantageous conditioning, an example of which is respondents' learning to answer filter questions negatively to avoid follow-up questions, decreases data quality.

These two types describe changes in reporting. A different type of panel conditioning is an actual change in attitudes, behavior or knowledge as a result of being interviewed. Such conditioning is disadvantageous since it introduces systematic bias into the sample because respondents would give other answers were there no previous measurement.

According to Bartels (1999) and Halpern-Manners and Warren (2012), panel conditioning is more likely when there is little time between surveys. Since online surveying is usually done at a higher frequency than face-to-face and telephone panel studies, there is a reason to be concerned about conditioning in online panels.

The goal of this chapter is to investigate panel conditioning in an online panel. It seeks to answer the questions of whether online panel experience generates learning and if learning is present, whether is it advantageous or disadvantageous for data quality.

5.2 Background

Since panel studies were introduced for measuring public opinion several decades ago, there has been a concern as to whether repeated interviewing affects respondents' opinions (Lazarsfeld, 1940). This concern has spread to changes in the reporting of attitudes and behavior (Kalton et al., 1989). Whether respondents change their reporting strategies or their actual opinions and behavior, the survey results differ from those that would have been obtained from the population originally represented by the study.

The emergence of online panels in recent years has led to an increase of studies on panel conditioning in the online mode. These studies do not provide straightforward conclusions about the phenomenon. Sometimes no evidence of panel conditioning is found (e.g., Axinn, Jennings, & Couper, 2013; Clinton, 2001; Dennis, 2001), and most times it is limited to certain types of questions (e.g., Binswanger, Schunk, & Toepoel, 2013; Kruse et al., 2009; Toepoel, Das, & van Soest, 2009). The studies usually focus on disadvantageous conditioning.

Findings from traditional modes are also unsystematic and sometimes contradictory (see reviews by Cantor, 2008; Sturgis, Allum, & Brunton-Smith, 2009; Warren & Halpern-Manners, 2012; Waterton & Livesley, 1989). The literature lacks definitive conclusions on the presence and magnitude of panel conditioning because of the absence of a unified theoretical framework (Sturgis et al., 2009; Warren & Halpern-Manners, 2012; Yan & Eckman, 2012) and suboptimal study designs with few experiments or quasi-experiments (Halpern-Manners & Warren, 2012; Torche, Warren, Halpern-Manners, & Valenzuela, 2012). The review of studies on panel conditioning according to their designs is provided in Appendix F. The remaining part of this section focuses on theoretical explanations of panel conditioning.

Panel conditioning results from two distinct processes: (1) learning the rules that govern the interview and (2) learning the content of the questionnaires. Both types of learning can be advantageous or disadvantageous for data quality.

Advantageous panel conditioning resulting from *learning the rules that govern the interview* manifests itself in two instances. First, respondents may provide more truthful answers. Participation in multiple interviews increases familiarity with the survey organization and induces respondents' trust. Respondents begin to feel more at ease and are willing to reveal sensitive or unflattering information. Waterton and Livesley (1989) find increased reporting of racial prejudice and fewer refusals to answer income questions, Bailar (1975) finds an increased willingness to admit to the socially undesirable status of not-working or being employed in an underground economy. Second, respondents learn that their attitudinal orientation is interesting to researchers, even if it is not well-formulated. This encourages respondents to choose a substantive answer over a "don't know"-answer (Binswanger et al., 2013; Waterton & Livesley, 1989).

On the other hand, becoming familiar with the surveying process may prompt respondents to engage in strategic behavior to reduce the burden of responding – *disadvantageous* panel conditioning. In previous interviews, respondents learn to recognize the type of questions that have dependent questions if answered in a certain way. Thus, if respondents want to reduce the burden of the follow-up questions they answer filter questions negatively and thereby avoid having to answer a series of dependent questions. Compared with first-time respondents, panel respondents report fewer functional limitations (Mathiowetz & Liar, 1994), hold fewer jobs, have fewer household members, are less likely members of political parties (Warren & Halpern-Manners, 2012), and are less likely to report using toothpaste (Nancarrow & Cartwright, 2007). All these questions were followed by dependent questions.

Why panel conditioning can lead to opposite effects is explained by the notion of "good" and "bad" respondents (Waterton & Livesley, 1989; see also Yan & Eckman, 2012). This dichotomy reflects respondents' motivation: as rational actors, respondents employ various strategies for the cognitively demanding task of answering survey questions. According to Tourangeau, Rips, and Rasinski (2000, p. 17), the processes in which respondents engage while producing an answer, vary in depth and quality of thoughts. The "high road" of arriving at an answer is a conscientious process of carefully going through the steps of comprehending the question, retrieving information, generating judgment, and reporting the answer. The "low road" is a superficial method to provide an answer without giving it much thought guided by motives other than accuracy. In terms of Krosnick and Alwin (1987), the opposing strategies used by respondents are optimizing and satisficing. Optimizing respondents seek optimal answers, while satisfiers select acceptable answers to minimize the psychological costs of seeking and providing accurate answers. Which strategy is used depends on the difficulty of the task, the respondents' ability, and their motivation (Krosnick, 1991). As respondents become more fatigued during the interview, they may shift their strategy from optimizing to satisficing. For web surveys, Galesic and Bosnjak (2009) show that in longer and thus more burdensome questionnaires, response quality decreases toward the end of the survey. This means that the notion of "good" or "bad" respondents is insufficient and question characteristics influencing respondent burden need to be considered. Two aspects of burden seem relevant: respondent's effort and respondent's stress – the amount of discomfort a respondent has to cope with – which is associated with the level of topic sensitivity (Bradburn, 1978). Furthermore, in panel surveys, the effort required to answer sensitive questions may be less than the effort for answering monotonous (boring) follow-up questions after a filter question. According to Bradburn (1978), "burdensomeness" is not an objective characteristic of a question and more difficult questions may be perceived as more challenging, which can explain why respondents provide substantive answers instead of a "don't know," since providing a substantive answer requires more cognitive effort than answering simple filter questions.

An alternative approach to explain panel conditioning is through *learning the questionnaire content*. This can also be advantageous and disadvantageous for data quality. *Advantageous* conditioning manifests itself in several ways. First, respondents provide more accurate responses because they know what is expected of them or prepare for future interviews (Kalton & Citro, 1993). Studies with validation data show that experience within the panel leads to more accurate reporting of income (Frick, Goebel, Schechtman, Wagner, & Yitzhaki, 2004; Rendtel, Nordberg, Jäntti, Hanisch, & Basic, 2004) and of receiving unemployment benefits (Yan & Eckman, 2012). Second, for sensitive questions, learning the content of the questions lessens respondents' threat and results in more honest responses. When asked a sensitive question for the first time, respondents experience initial shock and may lie in order to appear in a positive light. The second time, respondents have the experience of providing sensitive information without unfavorable consequences and answer honestly. According to Uhrig (2012), who finds heavier women reporting increased body weight with fewer rounded answers and item nonresponse in the British Household Panel Innovation Study, it is the familiarity with the questions and not the familiarity with the survey procedures that reduces social desirable responding. Third, learning the questionnaire content stimulates respondents to think about the interview topic or discuss it with others, which leads to more stable responses (Sturgis et al., 2009). Thinking more closely about their views motivates respondents to generate attitudes through rational judgment as opposed to reporting attitudes formed on the spot. To the effect of deliberation Sturgis et al. (2009) attribute lower rates of "don't know's" over the panel waves: respondents seek information on the topic, which results in increased "opinionation" and thus more substantive answers instead of "don't know's." Advantageous panel conditioning may also be explained by respondents' motivation.

Disadvantageous panel conditioning through learning the questionnaire content occurs in two cases: mere exposure to the questions and deliberation on the topic leading to attitude and/or behavior changes. Exposure to some questions increases respondent awareness and leads to differences in knowledge levels between experienced and first-time respondents (Coombs, 1973; Nancarrow & Cartwright, 2007; Toepoel et al., 2009; Yan, Datta, & Hepburn, 2011). Deliberation on the topic of the interview may lead to changes in attitudes or behavior. These changes are disadvantageous for data quality as they no longer represent the attitudes and behavior of non-interviewed counterparts. These effects were found in various domains: interviewing on the topic of cancer leads to rating good health as important (Bridge et al., 1977), being interviewed about vaccination increases reported vaccination levels (Battaglia, Zell, & Ching, 1996), pre-election surveys increase voter turnout (Clausen, 1968; Kraut & McConahay, 1973; Traugott & Katosh, 1979; Yalch, 1976), and interviewing recipients of unemployment benefits makes them more likely to move off of these benefits (Yan & Eckman, 2012). As opposed to other types of conditioning mentioned in this section, in disadvantageous panel conditioning due to learning the questionnaire content, respondents' greater motivation leads to adverse effects on data quality.

To summarize, panel conditioning may be explained by different mechanisms. Two findings of decreased "don't know" responses and decreased social desirability are attributed to opposing mechanisms of learning the interview procedures and learning the questionnaire content. The difficulties of attributing the results to either process may result from using study designs that, except experimental designs, do not allow for separating conditioning due to learning the surveying process from conditioning due to learning the questionnaire content.

Several analytic strategies have been employed to study panel conditioning. These strategies vary in their ability to tackle different confounding factors. The first option is to compare panel data to a fresh cross-section (e.g., Corder & Horvitz, 1989; Kruse et al., 2009; Waterton & Livesley, 1989) or to fresh respondents from another panel (e.g., Toepoel, Das, & van Soest, 2009). Provided that question wordings are the same in two studies for comparison, differential nonresponse and attrition may still be a confounding factor in both cases.

The second option is a rotating panel design: fresh groups of respondents are added to the panel at some specified intervals, that is, at each wave, and then compared to the panel component (e.g., Bailar, 1989; Pennell & Lepkowski, 1992; Silberstein & Jacobs, 1989). The advantage of this design is that question wording, the sponsor of the survey, the recruitment procedure, and the sampling frame are the same, however, the problem of nonresponse still holds.

The third option is to compare the groups with various tenure lengths within a panel (e.g., Clinton, 2001; Coombs, 1973; Dennis, 2001). This option resembles the rotating panel design since fresh groups added to the panel are compared with those that have participated longer. The difference is that the comparison here is not between a cross-sectional rotating group and the panel group, but that tenure within the panel is controlled for in the analysis, thereby providing several groups for comparison.

The fourth option is an analytic strategy rather than a specific survey design: testing hypotheses based on a theoretical model on one panel with no rotation groups (e.g., Sturgis et al., 2009; Van der Zouwen & Van Tilburg, 2001; Yan & Copeland, 2010). Sturgis et al. (2009) employ this analytical strategy when looking at the reliability and correlation of responses and endorsement of the "don't know" category across several waves. Apart from attrition, which is problematic for all designs described above, this design also is confounded with the true change.

The fifth and most favorable option is a true experimental design. In contrast to the design options described above, experimental studies are designed explicitly to study panel conditioning. A randomized experimental design in which the experimental and control groups receive different questionnaire content in the first wave of the panel and the same content – unchanged for the experimental group – in the second wave was used in early research on the possibility of changing behavior through interviewing (e.g. Bridge et al. (1977), Kraut and McConahay (1973); Traugott and Katosh (1979); Yalch (1976). Despite the recent increase in the usage of experiments to study panel conditioning, experimental studies based on large-scale surveys are still rare (but see Axinn et al., 2013; Torche et al., 2012; Uhrig, 2012). If experimental and control groups are asked similar questions, selective dropout due to the differences in questionnaire content (e.g., if the experimental and control group questions differ on the levels of burden or social desirability) becomes less likely, which means that attrition is not as problematic as with other designs.

Using experimental design, this chapter studies panel conditioning caused by learning the rules of the interview. The empirical study consists of two field experiments designed to investigate panel conditioning in a probability-based online panel. It includes a manipulation of rotating the questionnaire content and focuses on both advantageous and disadvantageous panel conditioning.

5.3 Two field experiments in an online panel

5.3.1 Data and method

The GESIS Online Panel Pilot (GOPP) is a probability-based online panel of Internet users, a methodological project at GESIS – Leibniz-Institute for the Social Sciences. Respondents were recruited by telephone using the dual-frame approach, which included landline and mobile numbers (more information on recruitment in Struminskaya, Kaczmirek, Schaurer, & Bandilla, 2014). The target population were German-speaking adults living in private households in Germany, who used the Internet for non-work-related purposes. At the end of the recruitment interview respondents were asked to provide their email addresses to join the panel and complete monthly surveys of 10–15 minutes. The overall response rate (AAPOR RR3) for the telephone interviews was 17.8%, with 4840 interviews conducted. With an Internet usage rate of 72.6%, 1665 of 3514 Internet users provided a valid email address to be sent an invitation to the online survey. The first online survey was started by 1010 respondents and completed by 934 persons. Recruitment rate – the proportion of respondents who provided initial consent over all eligible respondents – was 9%.

The recruitment fieldwork was done in three parts: February –April 2011, June–August 2011, and July–August 2011. Each monthly questionnaire had a leading topic. These features shape the design of the two studies, which employ changing the order (rotation) of the questionnaires (Table 5.1). The first row of Table 5.1 shows the “normal” order of online questionnaires one (Q1) to eight (Q8). In both experiments, the conventional order of the questionnaires was changed

Table 5.1. Design of two field experiments.

Date (approx.)	Mar 2011	Apr 2011	May 2011	Jun 2011	Jul 2011	Aug 2011	Sep 2011	Oct 2011	Nov 2011	Dec 2011	Jan 2012	Feb 2012
Sample 1	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8				
Sample 2				Q1	Q5	Q2	Q3	Q4	Q6	Q7	Q8	
Sample 3a					Q1	Q6	Q2	Q3	Q4	Q5	Q7	Q8
Sample 3b					Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8

Note: the markings indicate the questionnaires whose order was changed for the first (Q5) and for the second (Q6) field experiment.

so that the fifth questionnaire (first experiment) or the sixth questionnaire (second experiment) followed the first introductory questionnaire. The topic of questionnaire Q5 was ecology and the environment, and questionnaire Q6 was devoted to social networks.

Some clarification is needed on the nature of the experiments we conducted. An experiment is a study in which an intervention is deliberately introduced to observe its effects (Shadish, Cook, & Campbell, 2002). In randomized experiments, units are assigned to the treatment condition by a random process. In quasi-experiments, the assignment takes place through self-selection or administrator-selection (Shadish et al., 2002, p. 12). Randomized experiments can be performed in laboratory settings (lab experiments) or in real-world settings (field experiments). Field experiments have a higher external validity, which comes at a price of less control by the researcher than in laboratory experiments. In field experiments, although the randomization is performed in advance, selective nonresponse may threaten the randomization. For our manipulations, we use the term field experiment, because although telephone numbers drawn from the sample frame were randomly assigned to the recruitment groups, two logistical conditions prevent from treating the studies as controlled randomized experiments. First, the usage of differential incentives (an experimentally varied amount of 5 and 10 Euros with an additional 20-Euro reward in sample 1 vs. 0, 2, and 5-Euro conditions in samples 2, 3a, and 3b) might have caused self-selection. The second issue is differential attrition that might have occurred before the questionnaire rotation.

5.3.2 Hypotheses, measures, and analysis strategy

The first field experiment studies advantageous conditioning. Both types of it, namely respondents' choice of "don't know" answers and social desirability reduction, are addressed. It is hypothesized that through repeated interviewing, the survey situation becomes familiar and that respondents become more trusting and provide less socially desirable answers and fewer "don't knows." Waterton and Livesley (1989) view the "don't know" option as giving respondents less discomfort over a weak attitude. Others have argued that there is no reason why "don't know's" would be considered inappropriate by respondents (Sturgis et al., 2009). To test the hypothesis of panel conditioning due to decreased social desirability, questions in which "don't know" means admitting one's knowledge deficiency and may therefore be perceived as unflattering are used.

Hypothesis 1: More experienced respondents provide *more* “don’t know” answers in knowledge questions.

To test this hypothesis, knowledge items that focused on nuclear power plants were implemented in the questionnaire about ecology and environmental behavior. The items with response options “true,” “false,” and “don’t know” were replicated from the 2009 Eurobarometer 72.2:[19]

a. *The EU has the largest number of commercial nuclear power stations (for electricity production) in the world.*
b. *Nuclear power plants are the only producers of radioactive waste.*
c. *About a third of the electricity produced within the EU is produced by nuclear power plants.*
d. *New nuclear power plants are presently being constructed in Germany at this very moment.*

The first hypothesis predicts that respondents choose more as opposed to fewer “don’t know’s” which was found in previous studies, because “don’t know” represents a less desirable substantial answer option for knowledge questions. At the same time, “don’t know” is a no-opinion option, the choice of which by the respondent is one form of satisficing (Krosnick, 1991). Endorsement of this option may be caused not by more honest reporting, but rather by a lack of motivation. To test whether the endorsement of “don’t know” answers is due to satisficing, we use paradata, that is, data automatically gathered during the (computer-assisted) interview (Couper, 1998; Kaczmirek, 2009; Kreuter, 2013). The indicator of respondents’ motivation is response latency: less-motivated respondents would speed through the questionnaire.

Another explanation for Hypothesis 1 is respondents’ learning to endorse an explicitly offered “don’t know” in general. For that reason, respondents’ choice of the “don’t know” option across the questionnaire is analyzed. Generally in GOPP, no no-opinion options were used, but it was offered explicitly when it was a substantial answer choice or in questions replicated from self-administered surveys with a no-opinion option.

The second hypothesis is also derived from the idea of frequent interviews leading to social desirability reduction. The experiments are implemented in a discontinuous panel, that is, a panel in which question content does not repeat itself, so decreases of social desirability can be attributed to learning the survey procedure because learning the content was impossible as it was presented to both groups of respondents for the first time.

Hypothesis 2: More experienced respondents are less likely to provide socially desirable answers.

Measures for the second hypothesis were replicated from the International Social Survey Programme 2010[20] and focused on environmental behavior. Studies have revealed that extremely sensitive questions are less affected by panel conditioning than moderate ones (Uhrig, 2012), so the items for our study needed to be prone to social desirable responding,

[19] http://gesis.org/?id=2632&tx_eurobaromater_pi1[vol]=2632&tx_eurobaromater_pi1[pos1]=796

[20] http://gesis.org/issp/issp-modules-profiles/environment/2010/

but not be too extreme. Reporting on environmental behavior in Germany was shown to be prone to social desirability (Blasius, 1998; Neugebauer, 2004). Nevertheless, those items cannot be considered extreme and are therefore well-suited to our study. The items are:

How often do you…
a. …make a special effort to sort glass or tins or plastic or newspapers and so on for recycling?
b. …make a special effort to buy fruit and vegetables grown without pesticides or chemicals?
c. …cut back on driving a car for environmental reasons?
d. …reduce the energy or fuel you use at home for environmental reasons?
e. …choose to save or re-use water for environmental reasons?
f. …avoid buying certain products for environmental reasons?

The response options were: "always," "often," "sometimes," "never," and "doesn't apply" (for a and c). Experienced respondents are expected to show a higher endorsement of "never" and "sometimes." Responses to single items and the overall distributions of answers are compared. To study overall endorsement, the following index was constructed: $\left(\frac{\text{response to items a to f}}{\text{N}_{\text{answered items}}}\right) \times 6$. Furthermore, to investigate whether all of the items of the question battery on environment represent one dimension, a factor analysis was performed.

The second experiment aimed to capture disadvantageous conditioning. It is hypothesized that with increased fatigue, respondents use their knowledge about survey procedures acquired during previous interviews to reduce the burden. The evidence from the studies on disadvantageous panel conditioning (in section 5.2) suggests that disadvantageous conditioning was found in questions that had dependent questions and were thereby burdensome. Thus, our hypotheses for disadvantageous conditioning are:

Hypothesis 3: More experienced respondents foresee questionnaire loops (i.e., several rounds of follow-up questions) and try to reduce the burden of follow-up questions.

Hypothesis 4: More experienced respondents provide data of lower quality generating more item nonresponse and break-off.

In order to test these hypotheses, the names generator question was used. The names generator is a type of question used to collect information on respondents' personal social networks, including the names of persons with whom the respondent has a certain type of relationship indicated by the researcher, such as a friend, a relative, etc., and the information about these persons. The names generator asked for names or initials of five of the respondents' friends and included follow-up questions about every named friend. The topic and the question are burdensome: for ego-centered networks, Matzat and Snijders (2010) find lower data quality in a web survey than in a face-to-face survey. They show higher dropout during the network part of the survey, providing fewer names and more item nonresponse in questions that asked about the relationships between the network members.

The names generator is adapted from ALLBUS 2010 (Figure D1 in Appendix D). The follow-up questions on gender, age, relation to respondent, educational level, employment status, occupation, Germany being country of birth, country of birth (if not born in

Germany), citizenship, distance to the respondent, economic situation compared to the respondent's, and frequency of contact are also replicated from ALLBUS 2010.

Respondents who have been within the panel longer are expected to foresee repeated elements of the questionnaire (loops) because they have already experienced extensive filtering in the previous waves, especially in questionnaire Q2 on employment and questionnaire Q3 on household and family life. It is further expected that even if respondents did not recognize that questions would follow on the names generator screen, they would understand that questions will follow on every named person after seeing the next screen with the question *"Is [name of the first friend] male or female?"*. Respondents would then go back and change the provided number of friends. Experienced respondents are expected to do so at a higher rate. Apart from the comparisons of the distributions of the number of friends' names, page visits and keystroke information are used to detect whether respondents reduced the originally provided number of names.

Furthermore, experienced respondents are expected to provide less information about friends they named, either by avoiding answering the questions (item nonresponse) or by abandoning the survey (break-off) during the follow-up questions. In GOPP, edit-checks were implemented that prompted an answer if respondents clicked the submit button without answering the question. To proceed, they had to confirm the refusal to answer the question by clicking an additional checkbox (Figure D3 in Appendix D). Thus, initial item nonresponse might differ from the final data. Both initial and final item nonresponse are analyzed. One further indicator of data quality used in the analysis is response time.

5.3.3 The confounding factors

As noted in section 5.3.1, the main confounding factor of the field experiment is attrition. Several strategies have been used in studies on panel conditioning to correct for attrition. The first strategy is to exclude respondents from the unconditioned group who would attrite later in the course of the study (Waterton & Livesley, 1989). The second strategy is the implementation of more complicated models to decompose the total bias into panel conditioning and attrition effects by imposing additional assumptions on the attrition process (Das, Toepoel, & van Soest, 2011). The third strategy is weighting the observations of the analysis groups so that distributions of demographic variables match for "old" and "new" respondents and for responses provided for the first and second time (Sikkel & Hoogendoorn, 2008).

In our case, when changes in reporting are investigated within one panel wave, the straightforward method of excluding from analysis the persons of the control group who would attrite later is not a wise option since attrition may not be independent from the experiment. The second option is also not feasible because the design was not constructed to have first-time and second-time measures for the conditioned group. Weighting or matching the two groups therefore seems to be the best solution.

In order to account for attrition and to adjust for other design-related factors compromising the random assignment, propensity score weighting is used. Rosenbaum and Rubin (1983) define a propensity score as the conditional probability of receiving treatment T given the observed covariates X, or $e(x)=Pr(T=1|x)$. In the case of random assignment, propensity scores are known; when the treatment mechanism is confounded, propensity scores can be estimated from the data. For this purpose, a linear *logit* model is used: $log\left\{\frac{e(X_l)}{1-e(X_l)}\right\} = \alpha_0 + \beta_{1X_{l1}} + \beta_{2X_{l2}} + \varepsilon$, the fitted values $\hat{e}(X_l)$ are estimates of the propensity score (Rosenbaum, 2010, p. 166).

The choice of covariates for the specification model is crucial for the estimation of propensity scores. As the main purpose is to adjust for attrition, determinants of attrition should be included. In face-to-face studies, attrition is explained by respondent characteristics, attitudinal and situational characteristics, predictors of contact, survey design features, interaction with an interviewer, and data quality from previous interviews (Lepkowski & Couper, 2002; Watson & Wooden, 2009; Zabel, 1998). Predictors of attrition from face-to-face studies only partly apply to online studies. In the online probability-based LISS Panel, Lugtig (2014) finds age, education, income, being provided a computer, "big five" psychological characteristics, survey enjoyment, and burden to be predictive of attrition.

First, the groups are compared on demographic, attitudinal, survey experience and on survey-design-related variables. Second, propensity score weights are calculated. Normally, inverse propensity scores are used as weights (Rosenbaum, 1987). However, in our case, the goal is to correct for group differences and not to balance both groups to represent some population of interest. Therefore, following Schonlau, Van Soest, and Kapteyn (2007) and Hirano and Imbens (2001), the weights take the value of 1 for the conditioned group, and $\frac{\hat{e}}{1-\hat{e}}$ for the unconditioned group. Furthermore, the weights are normalized. Calculations using paradata as outcome variables are performed with the unweighted data. Although paradata are used for adjustment (e.g., Kreuter & Olson, 2013), there are no sufficient examples of the application of weighting to paradata.

5.4 Results

Since the experiments have a similar design, the results of the first field experiment (questionnaire 5) are presented followed by the results of the second field experiment (questionnaire 6). The fifth questionnaire was answered by 446 respondents from the two experimental groups. From the conditioned group, 273 started and completed the survey, resulting in no break-offs. From the unconditioned group, 176 respondents started the survey and 173 finished it, resulting in a break-off rate of 1.70%. The wave completion rate calculated as a proportion of respondents who started the survey over those who were invited was 66.42% for the conditioned group and 48.75% for the unconditioned group (Table 5.2). Differences in incentive conditions between the groups are a possible cause of the differences in completion rates. This may also explain the 6%-point difference in attrition calculated as a proportion of respondents who started the survey over the number of respondents who completed the first online survey Q1.

Table 5.2. Key measures for questionnaires 5 and 6 by experimental groups.

	Starting rate	Completion rate	Attrition	Cumulative RR[1]	N started	N completed	Break-off rate
Q5 conditioned	66.42%	66.42%	12.78%	5.98%	273	273	0.00%
Q5 unconditioned	48.75%	47.92%	18.52%	4.39%	176	173	1.70%
Q6 conditioned	65.28%	63.81%	16.61%	5.88%	267	261	2.25%
Q6 unconditioned	43.62%	38.07%	13.47%	3.93%	213	185	13.15%

[1](Recruitment rate × wave completion rate); RR stands for response rate.

5.4.1 Differences between the experimental groups

Table 5.3 presents demographic and attitudinal measures collected during recruitment and in the first online survey, as well as incentive conditions. Most of the variables were measured prior to the experiments. However, in two recruitment conditions, demographic variables were not measured in the telephone interview: (1) respondents refusing the telephone interview but willing to participate online could proceed to the recruitment; (2) the interview length was varied experimentally by excluding a number of questions. These questions were asked in questionnaires Q1 and Q2, which allows for solving the problem of missing data by using the values collected online.

For the unconditioned group, measures for the variables education and employment status were fielded after the experiment due to the questionnaire rotation. These characteristics are rather stable, so imputation from a subsequent wave seems unproblematic. The number of cases for which education and employment status are missing due to attrition after the experiment is negligible. Missing values for the previous survey experience were imputed with the information about online survey experience. Overall, the proportion of missing values due to nonresponse or break-off at the first questionnaire is 4.25%.

The only statistically significant variable by which the groups differ from one another is survey experience in the previous 12 months. This can either be caused by attrition or by initial inequality of the groups (e.g., higher incentives offered to the conditioned group). The percent of conditioned respondents who started the second online survey Q2, which would be equivalent to the value of the variable "survey experience" for unconditioned respondents reported in Table 5.3, is 36.40% (N=283). This indicates that the groups were unequal from the start of the panel, which can be corrected by propensity score weighting.

5.4.2 Propensity-score weighting to correct for group differences

In order to estimate propensity scores, it is advisable to include variables that differ significantly between the groups, important predictors of treatment and outcome, and mildly significant predictors of treatment (Schafer & Kang, 2008, p. 297). We included the variable of survey experience in the previous 12 months, by which the groups differ at a statistically significant level, and the incentive indicator for which significance tests could not be performed due to

Table 5.3. Descriptive statistics for experimental groups and differences between the groups.

	Unconditioned (T=0)			Conditioned (T=1)			Difference	
Variable	**M/%**	**SD**	**N**	**M/%**	**SD**	**N**	**Test statistic**	***p*-level**
Male	55.17%	0.499	174	52.38%	0.500	273	$\chi^2(1)=0.333$	0.564
Age	41.80	15.495	171	42.97	14.308	272	$t=-0.808$	0.419
Education		0.664	165		0.684	269	$\chi^2(2)=2.046$	0.359
low	9.70%			10.78%				
middle	23.03%			28.62%				
high	67.27%			60.59%				
In paid work	66.67%	0.473	168	70.74%	0.456	270	$\chi^2(1)=0.806$	0.369
Trust	2.74	0.656	163	2.78	0.647	263	$t=-0.609$	0.543
Political interest	2.72	0.731	163	2.65	0.775	265	$t=0.992$	0.322
Self-assessed health	3.72	0.899	163	3.68	0.925	265	$t=0.464$	0.643
Survey participation past 12 months	55.21%	0.499	163	38.49%	0.487	265	$\chi^2(1)=11.412$	0.001
Overall evaluation Q1	3.78	0.545	163	3.79	0.578	265	$t=-0.170$	0.866
Q1 interesting	89.44%	0.308	161	84.91%	0.359	265	$\chi^2(1)=1.778$	0.182
Q1 varied	87.65%	0.330	162	84.09%	0.366	264	$\chi^2(1)=1.025$	0.311
Q1 important f. science	81.05%	0.393	153	75.38%	0.432	260	$\chi^2(1)=1.769$	0.183
Q1 long	9.38%	0.292	160	10.65%	0.309	263	$\chi^2(1)=0.176$	0.675
Q1 difficult	2.50%	0.157	160	1.14%	0.106	264	$\chi^2(1)=1.141$	0.285
Q1 too personal	8.13%	0.274	160	5.68%	0.232	264	$\chi^2(1)=0.964$	0.326
Incentive		0.778	174		0.501	273	—	—
0	23.56%			—			—	—
2	37.36%			—			—	—
5/5+bonus	39.08%			51.65%			—	—
10+bonus	—			48.35%			—	—

Note: sample sizes vary due to missing values. M=mean, %=proportion, SD=standard deviation, N=sample size.

5

partial overlap of the conditions, but which is thought to influence response and attrition. We also included education as a proxy measure of cognitive ability that can influence answers on the knowledge items. Toepoel et al. (2009) found that education predicts panel conditioning in knowledge questions. Our data show that two knowledge items are significantly related to education (item 1: $\chi^2(4)=5.736$, $p=0.220$, item 2: $\chi^2(4)=22.231$, $p<0.001$; item 3: Fisher's exact=0.085; item 4: Fisher's exact<0.001). Thus, the model for estimation of propensity scores is:

$$\hat{e} = \alpha_0 + \beta_{1survey\ experience} + \beta_{2incentive} + \beta_{3education} + \varepsilon.$$

In order to match observations, the incentive variable, which only partly overlaps between the groups, is recoded in an indicator variable, which contrasts respondents who received 5 Euros against all other incentive conditions. After estimating the propensity scores, we checked whether the treated and control groups achieved the covariate balance. The most striking difference was observed in survey experience; therefore, this measure is compared using the weighted and the unweighted data. The unweighted proportions for the unconditioned and conditioned groups are 55.21% (Table 5.3) and 38.49%, respectively ($\chi^2(1)=11.412$, $p<0.001$). The weighted proportion for the unconditioned group is 36.46%, which is no longer significantly different from the conditioned group. The weighted proportion for the conditioned group equals its unweighted value because as specified in section 5.3.3, propensity score weights take the value of 1 for the conditioned group. The unweighted indicator variable for incentives for the unconditioned group takes the value of 39.08%, it is 51.65% for the conditioned group ($t=-2.611$ $p=0.009$), the weighted proportion is 53.51% for the unconditioned group. The weighted distribution of education for the unconditioned group is: 13.45% – low, 24.08% – middle, and 62.47% – high. The distribution of propensity sores for the conditioned group has the mean 0.635, standard deviation 0.090, minimum=0.468, and maximum=0.781. The unconditioned group has $M=0.595$, $SD=0.102$, minimum=0.468, and maximum=0.781 (Figure D2 in Appendix D). The ranges of estimated propensity scores match, and the overlap of the distributions of the propensity scores seems sufficient to proceed. Therefore, the analysis for the first field experiment on advantageous panel conditioning is based on the weighted data (sections 5.4.3 and 5.4.4). For the second field experiment on disadvantageous panel conditioning, a similar assessment of the group differences is undertaken (see section 5.4.5).

5

5.4.3 Advantageous panel conditioning: Answers to the knowledge question

Hypothesis 1 predicts that more experienced respondents provide more “don’t know” answers in knowledge questions. Table 5.4 shows the distributions of answers by experimental groups.

Across all knowledge items, conditioned respondents answer “don’t know” more frequently, with differences ranging from 0.99% to 9.47%. Two differences in distributions of “don’t know” answers and substantial answer options are statistically significant: for item 1 ($\chi^2(2)=10.984$, $p=0.004$) and for item 3 ($\chi^2(2)=7.959$, $p=0.019$). A comparison was performed with chi-squared tests for equality of distributions.

Conditioned and unconditioned respondents may differ in their response patterns for the whole question. An index of “don’t know” responses across all items of the knowledge question was constructed. The comparison of the groups is presented in Table 5.5.

Overall, the two distributions differ significantly ($\chi^2(2)=14.992$, $p=0.005$). The difference of nearly ten percentage points in no endorsement of “don’t know” suggests that less experienced respondents provide answers to all items of the knowledge question at a higher rate than do those respondents who have been with the panel longer.

Alternative explanation: Satisficing

The higher endorsement of “don’t know” by the conditioned group may be explained not by more honest responding, but by a higher level of satisficing. If respondents satisfice, they

Table 5.4. Distributions of answers to the knowledge questions by groups, propensity score weighted.

Items	Answer	Unconditioned %	Unconditioned SE	Conditioned %	Conditioned SE	Difference
(1) The EU has the largest number of commercial nuclear power stations (for electricity production) in the world.	Right	32.14	0.040	25.64	0.026	
	Wrong	30.81	0.040	27.84	0.027	
	Don't know	37.05	0.041	46.52	0.030	9.47***
(2) Nuclear power plants are the only producers of radioactive waste.	Right	15.44	0.030	16.12	0.022	
	Wrong	73.10	0.037	71.43	0.027	
	Don't know	11.46	0.026	12.45	0.020	0.99
(3) About a third of the electricity produced within the EU is produced by nuclear power plants.	Right	66.34	0.040	60.07	0.030	
	Wrong	14.33	0.030	13.92	0.021	
	Don't know	19.33	0.034	26.01	0.027	6.69**
(4) New nuclear power plants are presently being constructed in Germany at this very moment.	Right	2.97	0.013	4.40	0.012	
	Wrong	81.75	0.034	79.85	0.024	
	Don't know	15.28	0.032	15.75	0.022	0.47

Note: Unconditioned $N_{unweighted}$=174, $N_{weighted}$=271.459, conditioned N=273, linearized standard errors, **p<0.05, ***p<0.01. The right answers for the items can be found in Appendix D.

Table 5.5. Endorsement of "don't know" across all knowledge items.

Number of DK-answers	Unconditioned %	Unconditioned SE	Conditioned %	Conditioned SE
0	48.52	0.043	38.83	0.030
1	28.51	0.039	34.80	0.029
2	16.41	0.032	16.12	0.022
3	4.46	0.017	7.32	0.016
4	2.10	0.011	2.93	0.010

Note: unconditioned $N_{unweighted}$=174, $N_{weighted}$= 271.459, conditioned N=273, linearized standard errors.

should take less time to answer the question. Contrary to this expectation, however, the conditioned group needed longer to answer the knowledge question. The mean answering time was 45.424 seconds (*SD*=45.317) for the unconditioned group vs. 49.775 seconds (*SD*=55.976) for the conditioned group. The difference is not statistically significant. However, it is the first hint that allows for ruling out satisficing by more experienced panelists. The second key figure is the time needed to read the question and form the answer. Although experienced respondents take longer to provide answers, they may spend less time reading the question. The paradata script allows for measuring the time until the first click on a page, which is generally the endorsement of the answer to the first item. This indicator also includes the time taken to form the answer to the first item, but is still a good and the only proxy of the reading time obtainable without eye-tracking (Lenzner, Kaczmirek, & Galesic, 2011). For the unconditioned group, the time is 17.366 seconds (*SD*=16.551, *CI*=[14.875; 19.857]), for the conditioned group, it is 21.527 seconds (*SD*=42.035, *CI*=[16.518; 26.535]). The difference is again not statistically significant. There is a vast difference of standard deviations between the groups, which might be caused by experienced respondents' using search engines to find the right answers and thus needing more time. A proxy measure leaving the survey window is used to rule out this explanation. About 3.71% of respondents who completed the survey without a time break left the window. The number of times leaving the window ranged from 1–6 times with a mean value of 0.067 (*SD*=0.415) for the unconditioned group and *M*=0.075 (*SD*=0.493) for the conditioned group. The difference between the groups is not statistically significant. When those who leave the page are dropped from the analysis, still no significant difference between the groups in response latencies is found.

Alternative explanation: Overall higher "don't know" endorsement

Experienced respondents' higher endorsement of "don't know" and taking more time to answer the question could alternatively be explained by learning to use the explicitly offered no-opinion option. For that reason, choosing the "don't know" option across the entire questionnaire is analyzed. Overall, "don't know" was offered in 27 items in the fifth questionnaire, including the 4 knowledge items. The endorsement by respondents ranged from 0–17 times. On average, unconditioned respondents chose the "don't know" option 1.641 times (SE_{lin}=0.192) vs. a mean endorsement of 1.817 times (SE_{lin}=0.139) by conditioned respondents. The difference between the groups is not statistically significant. The comparison of "don't know" responses relative to the number of answered questions shows a similar pattern.

In sum, analyses for Hypothesis 1 suggest that there is some evidence that experienced respondents answer knowledge questions more honestly. Although single-item analyses display inconsistent results in significance differences, the direction of these differences is as predicted. Based on paradata analysis, experienced and inexperienced respondents do not differ in choosing the "don't know" option. Nevertheless, the results are in line with Waterton and Livesley (1989), who argue that respondents learn the rules of the interview in the course of several surveys. In our case, the learning manifests itself in the manner in which respondents have no difficulty revealing the no-opinion option.

5.4.4 Advantageous panel conditioning: Reports of environmental behavior

According to Hypothesis 2, more experienced respondents are less likely to provide socially desirable answers than are less experienced respondents. Table 5.6 shows the weighted distributions of environmental items for the unconditioned and conditioned groups. In the case of socially desirable responding, respondents from the conditioned group would choose the answer options "never" and "sometimes" at a higher rate than respondents from the unconditioned group. However, the distributions of answers do not differ significantly. Respondents with more experience within the panel do not endorse the categories "never" and "sometimes" more often

Table 5.6. Descriptive statistics on environmental behavior by experimental groups (%)

Answers	Item a	Item b	Item c	Item d	Item e	Item f
Unconditioned						
always	73.77 (0.036)	13.79 (0.028)	3.40 (0.016)	13.90 (0.029)	11.40 (0.027)	5.88 (0.022)
often	22.67 (0.034)	43.55 (0.042)	32.29 (0.040)	42.05 (0.041)	38.82 (0.041)	38.65 (0.041)
sometimes	3.02 (0.013)	33.06 (0.040)	44.72 (0.043)	31.85 (0.039)	33.81 (0.039)	43.43 (0.041)
never	0.54 (0.005)	9.60 (0.024)	19.59 (0.035)	12.20 (0.028)	15.97 (0.031)	12.04 (0.028)
Conditioned						
always	73.80 (0.026)	15.07 (0.022)	4.31 (0.012)	8.42 (0.017)	12.82 (0.020)	5.13 (0.013)
often	20.66 (0.025)	43.75 (0.030)	33.33 (0.030)	45.79 (0.030)	37.00 (0.029)	39.19 (0.030)
sometimes	4.80 (0.013)	29.04 (0.028)	43.14 (0.031)	34.07 (0.029)	33.70 (0.029)	40.66 (0.030)
never	0.74 (0.005)	12.13 (0.020)	19.22 (0.025)	11.72 (0.019)	16.48 (0.022)	15.02 (0.022)
$\chi^2(3)$	3.532	3.472	0.868	7.268	0.766	2.783
p	0.317	0.324	0.833	0.064	0.858	0.426

Note: absolute sample size varies due to the "does not apply" category for unconditioned between 161 and 173, and between 255 and 273 for conditioned, weighted N for unconditioned between 252.766 and 274.261; linearized SE in parentheses. No environmental items are significantly related to education; therefore, it is not used to construct the propensity score weights.

than inexperienced respondents. A factor analysis performed with both the GOPP data and the ISSP data confirms that all items load on one factor (details in Appendix D).

The comparison of the indexes for all environmental items shows no significant differences between the groups: for the unconditioned group, the mean is 15.939 (linearized SE=0.272); for the conditioned group, the mean is 15.849 (linearized SE=0.201, t=0.27, p=0.791). Taken together, the data show no support for the hypothesis that more experienced respondents give less socially desirable answers.

5.4.5 Disadvantageous panel conditioning: The names generator question

The remaining results sections report findings of the second field experiment. The questionnaire Q6 was started by 480 respondents and finished by 446 respondents (Table 5.2). The break-off rate was higher in the unconditioned group (13.15% vs. 2.25%). Prior to reporting the results, we investigate whether there are differences between the experimental groups caused by attrition and other confounding factors that need to be corrected using propensity score weighting. The procedure of comparing the two groups is the same as for the first field experiment. The groups are compared on the same set of demographic and attitudinal variables as described in section 5.4.1. None of the examined variables show significant differences (Table D3 in Appendix D). Additionally, the relationship of the outcome variable (number of friends reported) and general trust, for which the dependence might be theoretically expected, was examined. These variables did not show sufficient correlation (r=0.072). None of the variables show significant differences, so it is unnecessary to perform propensity score weighting. The results presented in this section and in section 5.4.6 are based on the unweighted data.

Hypothesis 3 predicted that conditioned respondents would provide fewer names. Table 5.7 shows the distributions of the number of the provided names by experimental groups. In both groups, the majority of respondents provide five names (Table 5.7). Somewhat more respondents from the conditioned group provide no names; however, the difference is negligible. Neither the overall pattern of providing the names, nor the mean number of names differ significantly between the groups (Fischer exact=0.714, t=0.512, p=0.304).

Respondents were predicted to change their answers after having seen the first follow-up question. However, distributions based on the final data (Table 5.7) do not differ substantially from distributions of initially provided names. Overall, 44 persons visited the page with the names generator more than once. Of those persons, 8 provided less than five names in the final data: 4 from each experimental group. Curiously, respondents who provided five names and visited the names generator more than once, changed the names, the order of friends, or nothing about the names (Kaczmirek, 2013). Only two respondents reduced the number of names they originally provided. Thus, the expectation that respondents with more experience within the panel would provide fewer names is not confirmed.

5.4.6 Disadvantageous panel conditioning: Item nonresponse, break-off, and speeding

In order to assess the quality of data provided by conditioned and unconditioned groups, the following indicators are used: break-off and item nonresponse at the names generator and in the follow-up questions, and answering time. As Table 5.2 showed, break-off rates differ substantially between the conditioned and the unconditioned groups. When comparing break-off rates, not counting those abandoning the survey prior to the names generator, the difference between the groups is still significant (7.51% unconditioned and 1.12%

Table 5.7. Number of friends' names by groups.

Number of friends´ names	Unconditioned %	Conditioned %
0	5.58	6.82
1	2.54	1.89
2	1.52	3.41
3	6.60	6.82
4	11.17	8.33
5	72.59	72.73
Sample size	197	264
Mean	4.330 (1.369)	4.261 (1.460)

Note: standard deviations of the means in parentheses.

conditioned, Fisher exact=0.001). The small group sizes of those who broke off do not allow for a more detailed investigation on the basis of which sound conclusions can be reached.

Groups did not differ significantly on item nonresponse to the name generator: 6.82% for conditioned und 5.58% for unconditioned ($\chi^2(1)=0.292$, $p=0.589$). For those who provided at least one name and did not break-off during the follow-up questions, the difference between the groups in item nonresponse for follow-up questions is negligible. About 97.95% of conditioned respondents and 98.29% of unconditioned respondents provided all information about every friend they named. Item nonresponse to the names generator, including those who wanted to skip the question but provided an answer and those who skipped the question anyway, did not differ between the groups (unconditioned: 6.09%, conditioned: 8.71%, $\chi^2(1)=1.104$, $p=0.293$). Item nonresponse for the follow-up questions before edit-checks for respondents who provided at least one name and did not break off during the network block was 10.25% for the conditioned and 9.14% for the unconditioned, and this difference is not significant. The number of follow-up questions at which edit checks were activated ranged from 1–17 for the conditioned (M=2.12, SD=3.295) and 1–8 for the unconditioned group (M=1.625, SD=1.746), yielding no statistically significant difference. There is no indication that respondents with more experience sped through the questionnaire: mean duration for questionnaires completed with no interruption is 820.916 seconds for the unconditioned and 812.556 seconds for the conditioned group. The difference is not statistically significant.

In sum, little evidence of disadvantageous panel conditioning is found. Both Hypothesis 3 and Hypothesis 4 can be rejected. Contrary to our expectations, respondents with more experience within the panel do not name fewer personal network members to avoid the burden of the follow-up questions than do respondents from the less experienced group. Based on break-off and item-nonresponse during the network section of the online questionnaire, experienced respondents do not differ from less experienced respondents. Furthermore, there is no evidence that more experienced respondents speed through the questionnaire. Both respondent groups appear to be very motivated, with over 95% of respondents providing full information about the persons they named.

5.5 Conclusion and discussion

Panel conditioning may occur due to respondents' learning the content of questionnaires or becoming familiar with the surveying process. This chapter studies the latter process using data from two field experiments in a probability-based online panel. Both experiments studied changes in reporting and involved rotation of the questionnaire content. It is argued that the discussion that sees panel conditioning as a source of nonsampling errors in panel surveys is not comprehensive: panel conditioning can be advantageous and disadvantageous to data quality. Both types of conditioning are studied.

The first experiment tested advantageous panel conditioning. It was hypothesized that respondents more experienced within the panel provide more honest responses of "don't know" to the knowledge questions and less socially desirable responses on environmental behavior. Overall, those hypotheses find limited support. Significant differences with more experienced respondents providing more "don't' know" answers to the knowledge question were found in two out of four items. Experienced respondents did not provide less socially desirable answers about their environmental behavior.

The second experiment studied disadvantageous panel conditioning. Respondents were asked to provide up to five names at the names generator question, followed by extensive follow-up questions. There is no evidence of experienced respondents providing fewer names to reduce burden. Furthermore, no evidence was found that experienced respondents provided data of lower quality. Both groups appeared to be highly motivated, with over 95% of respondents providing five names or initials and over 95% providing full information on every person they named. In sum, the results suggest very limited panel conditioning in reporting behavior.

Due to several limitations, a certain level of caution is required when generalizing these results. First, the reason why we see some conditioning effect on two items of the knowledge question and none on the other two may have to do with the unequal levels the items' difficulty. If answers to some knowledge items are widely known, it is not possible to detect the conditioning effect. Future studies should account for the baseline knowledge when choosing the items. This baseline knowledge should be relatively low. Second, the environmental items, which showed no differences between conditioned and unconditioned respondents, might not have been sensitive enough to capture the effect. It has been argued that extremely sensitive items do not produce conditioning effects (Uhrig, 2012), so the optimal level of item sensitivity should be determined. It would be beneficial to investigate items with various levels of sensitivity within one study. Third, the names generator might have had too few name slots. Matzat and Snijders (2010) report a mean network size of 1.5 persons when 3 name slots were provided in an online survey. In the GESIS Online Panel Pilot, the question with 3 slots was tested with a group of pretest participants who were recruited in the same manner and received the questionnaires a few weeks prior to the actual panel members. In this pretest, which is not included in any of the analyses here, almost everybody provided 3 names, so we increased the number of slots to 5. It is advisable to further increase this number since the majority of respondents seem to provide as many names as they are asked for. An increase in the number of slots for names might lead to different results.

The results of the second field experiment might also reflect that the names generator was a new type of question for both experienced and less experienced respondents. Kreuter, McCulloch, Presser, and Tourangeau (2011) find that in a questionnaire with multiple similar blocks with filter questions followed by extensive follow-ups, respondents "forget" their burdensome experience with the introduction of the next block, which has another topic. The solution would be to implement a similar-looking question on another topic in the wave prior to the experiment, so that the experienced (conditioned) group becomes familiar with this particular type of question.

Despite these limitations, this chapter reaches a valuable conclusion for data users from online panels: panel conditioning is not a threat to data users' conclusions when the cross-sectional questions prevail in their questionnaires (i.e., when the panels are discontinuous). Moreover, data quality might improve over the waves. The effects found for advantageous panel conditioning are small (effect sizes<0.1using formulas of Cohen, 1992), but no disadvantageous conditioning was found in any case.

What are the implications of this study for researchers studying panel conditioning and for survey practitioners? First, the study provides an experimental design that can be implemented in discontinuous and in "true" longitudinal panels with repeated measures. It addresses handling the confounding factors, a challenge which every study on panel conditioning will face. Experimental designs allow for reducing the number of such factors, but attrition remains problematic. This chapter suggests solving the problem of attrition with propensity score weighting. Second, this study advances the field by focusing on advantageous and disadvantageous conditioning and by isolating panel conditioning due to learning the survey procedure from panel conditioning due to learning the questionnaire content. Similar findings on "don't know" response patterns and social desirability reduction have been attributed to those opposing mechanisms. This study provides reason to believe that such panel conditioning is less likely to be due to learning the interview procedure. Future analyses aimed at exploring the role of the content of the questions should take question characteristics (sensitivity, offering a no-opinion option, monotonous follow-up questions) into account. Finally, the analyses in this chapter rely heavily on paradata. This approach is unique for panel conditioning studies to date, even for those studies that are carried out in the online mode. Response latencies, moving across the questionnaire, keystroke information, and other indicators are used to make inferences about respondents' motivation, which may change during the survey and depending on question characteristics. Studies on panel conditioning do not rarely have a low number of cases because large-scale panels are reluctant to implement conditioning experiments (e.g., here achieved power for study components varied between 0.09 and 0.29). In such studies, if no panel conditioning is found or in order to study subgroup differences, paradata may be used for robustness checks. Overall, paradata allows for ruling out alternative explanations as has been demonstrated by this study, and they should be paid more attention in future research on panel conditioning.

Chapter 6

Mode system effects in an online panel study: Comparison of the whole systems of data collection

This chapter was co-written with Edith de Leeuw and Lars Kaczmirek and is currently under review.

Abstract

One of the methods for evaluating online panels in terms of data quality is comparing the estimates that the panels provide with benchmark sources. For probability-based online panels, high-quality surveys or government statistics can be used as references. If differences among the benchmark and the online panel estimates are found, these can have several causes. First, the question wordings may differ between the sources, which may lead to differences in measurement. Second, the reference and the online panel may not be comparable in terms of the sample composition. Finally, since the reference estimates are usually collected face-to-face or by telephone, mode effects might be expected. In this chapter, we investigate mode system effects, an alternative to mode effects that does not focus on measurement differences between the modes but also incorporates survey design features into the comparison. The data from a probability-based offline-recruited online panel is compared to the data from two face-to-face surveys with almost identical recruitment protocols. In the analysis, the distinction is made between factual and attitudinal questions. The results show that the online panel differs from face-to-face surveys in both attitudinal and factual measures. However, the reference surveys only differ in attitudinal measures and show no significant differences for factual questions. We attribute this to the instability of attitudes and thus show the importance of using two surveys of the same mode for comparison. Along with significances, the effect sizes of the differences are reported. This allows for better comparability of the surveys.

6.1 Introduction

Several large-scale panel and repeated cross-sectional surveys incorporate or are planning to incorporate the online mode for data collection to reduce costs or to maximize contact and response rates. Some surveys begin as online panels with an interviewer-administered recruitment continued with a web-based data collection (e.g., the LISS Panel[21] in the Netherlands and the KnowledgePanel[22] of GfK Custom Research, formerly Knowledge Networks in the United States). Other surveys switch from the interviewer-administered to the online mode (e.g., the Netherlands Kinship Panel Study[23]), employ an additional online component (e.g., ANES 2008-2009 Panel Study[24]), or experiment with including the online mode along with the other modes (ESS experiments on mixing modes[25], Understanding Society Innovation Panel in the UK[26], and Labor Force Survey and Crime Victimization Survey in the Netherlands[27]).

Data users need to know that irrespective of the mode in which data were collected, it is possible to make valid inferences about the processes that data users study and that panel or trend data are comparable, that is, not influenced by the mode change. Ideally, to estimate a pure mode effect, that is, a "measurement bias that is specifically attributable to the mode" (Biemer & Lyberg, 2003, p. 207), a randomized experiment that assigns respondents to different modes would be conducted. However, logistic and budget constraints often prevent survey practitioners from doing so. An alternative is to measure a mode system effect, that is, to compare whole systems of data collection. The system is defined as an "entire data collection process designed around a specific mode" (Biemer & Lyberg, 2003, p. 208). According to this approach, in addition to the mode of data collection, the system of data collection includes other design parameters and procedures, such as sampling, interviewer-related factors, questionnaire contents, etc.

In this chapter, we study the mode system effects of three data collection systems, comparing the data from an online panel to two face-to-face reference surveys. We concentrate on two types of questions: factual and attitudinal questions. Questions about attitudes and opinions place different demands on respondents' cognitive processes than do factual and behavioral questions. Attitude is defined as a collection of feelings, beliefs, and knowledge about an issue – considerations that have different levels of accessibility. While forming an answer, respondents process these considerations, which requires deliberation and effort (Tourangeau, Rips, & Rasinski, 2000, pp. 179-180). In a face-to-face survey, an interviewer initiates the interaction and thereby controls the pace of the interview (de Leeuw,

21 http://www.lissdata.nl/lissdata/

22 http://www.knowledgenetworks.com/knpanel/

23 http://www.nkps.nl/NKPSEN/nkps.htm; see also Hox, de Leeuw, Klausch, and Zijlmans (2014).

24 http://www.electionstudies.org/studypages/2008_2009panel/anes2008_2009panel.htm

25 http://www.europeansocialsurvey.org/methodology/mixed_mode_data_collection.html

26 https://www.understandingsociety.ac.uk/about/innovation-panel; see also Auspurg et al. (2013)

27 See Buelens et al. (2012)

2005), so the time might not be sufficient for a respondent to process the considerations needed for generating responses to an attitudinal question. In a self-administered online survey, by contrast, it is the respondent who controls the survey situation, including the pacing, which allows for taking more time if needed to answer an attitudinal question. We hypothesize the mode system effects to be more pronounced in attitudinal questions. In order to investigate the mode system effect, we (1) control for differences in sample composition and (2) concentrate on questions with the same question wordings, because these two factors may be alternative causes of differences in estimates (de Leeuw & Hox, 2011).

6.2 Data

We use data from the GESIS Online Panel Pilot[28], a probability-based online panel of Internet users in Germany, and two cross-sections from the German General Social Survey as reference surveys.

GESIS Online Panel Pilot

The GESIS Online Panel Pilot (GOPP) is a telephone-recruited (landline and mobile)[29] online panel of Internet users aged 18 and older who live in private households in Germany. The recruitment was performed in three sequential study parts using almost identical recruitment protocols. The recruitment for the first part was done in February 2011, the recruitment for the second part was done in June-July 2011, and the third part was recruited in July-August 2011. After a short telephone interview, respondents were asked to provide their email addresses in order to join an online panel. Respondents who agreed would then be sent email invitations to online surveys of 10-15 minutes every month for eight months in total. Prospective panel members were offered incentives of 0, 2, 5, or 10 Euros, which were varied experimentally. An additional bonus of 20 Euros for completing all eight online questionnaires was offered (see Table 6.1 for more details). Every questionnaire had a leading topic and most questions were asked once. We use single questions from questionnaires one through five and seven.

German General Social Survey "ALLBUS"

The German General Social Survey ("ALLBUS") is a general population survey on attitudes, behavior, and social structure in Germany. It has been conducted by GESIS biannually since 1980. The survey is administered face-to-face. For our analyses, we use ALLBUS 2010[30] and ALLBUS 2012[31], which both implemented a two-stage disproportionate random sample of

28 https://dbk.gesis.org/dbksearch/sdesc2.asp?no=5582&db=e&doi=10.4232/1.11570

29 For details on telephone recruitment, see Chapter 2, Section 2.2.2 of this dissertation.

30 http://www.gesis.org/en/allbus/study-profiles/2010/

31 http://www.gesis.org/en/allbus/study-profiles/2012/

Table 6.1. Sample and survey design description for GOPP and ALLBUS surveys.

Survey characteristic	GOPP	ALLBUS 2010	ALLBUS 2012
Sample source	Adjusted RDD (landline+mobile)	Two-stage probability: municipalities and individuals from municipal registers	Two-stage probability: municipalities and individuals from municipal registers
Target population	Individuals living in Germany aged 18 and older, who use the Internet	Individuals living in households in Germany aged 18 and older	Individuals living in households in Germany aged 18 and older
Recruitment mode	CATI	CAPI	CAPI
Data collection mode	Online, CATI	CAPI	CAPI
Response/Participation rates	CATI AAPOR3: 17.8%; Panel participation rates per questionnaire: Q1: 54.92%; Q2: 46.89%; Q3: 45.96%; Q4: 44.43%; Q5: 44.08%; Q7: 42.09%	34.4%	37.6%
Fieldwork	February 2011–May 2012	May–November 2010	April–September 2012
Incentives	0, 2, 5, 10, bonus 20 EUR (experimental)	0, 10, 20 EUR (experimental)	10 EUR
N	1114 (started Q1-Q5,Q7)	2827	3480
N Internet users	1114	1869	2525
Percent Internet users[a]	71.5%	67.1%	73.6%

[a] Unweighted proportion for GOPP, design-weighted for ALLBUS 2010 and 2012.

individuals living in private households in Germany and aged 18 and older. Both ALLBUS 2010 and ALLBUS 2012 were administered by the same fieldwork agency that implemented similar procedures for contacting and interviewing the respondents. One difference is that in ALLBUS 2010, an incentive experiment was employed in which respondents could receive 10 Euros, 20 Euros, or no incentive (Wasmer, Scholz, Blohm, Walter, & Jutz, 2012, p. 51), whereas in ALLBUS 2012, all respondents were paid 10 Euros. The difference between the ALLBUS surveys and GOPP is that GOPP's target population only includes Internet users. A question on private Internet use is asked in both ALLBUS surveys, so for our analysis, we are able to compare GOPP with full samples of the ALLBUS surveys and with the subsamples of Internet users.

6.3 Measures

We use a selection of factual and attitudinal questions that were asked in ALLBUS 2010 and replicated in GOPP. ALLBUS has item batteries and single questions on opinions, which are repeated over the years in order to analyze social trends. The wording of the demographic questions generally does not change between the ALLBUS surveys. We use questions that were present in both ALLBUS 2010 and 2012 and replicated in the GOPP. The questions we use in the analysis were not repeatedly measured in GOPP but distributed over the questionnaires, so that each question that we use for comparison was asked once. This prevents possible confounding of the answers due to learning effects (panel conditioning). We include questions that had the same question wordings (see Appendix E for details) and were asked of the whole sample, that is, not preceded by filters.

Overall, 12 attitudinal and 7 factual items fit these criteria. Attitudinal items included respondents' assessment of the current economic situation in Germany and the economic situation in one year, the assessment of respondents' own financial situation and prospective financial situation in one year, general health, religiosity, self-assessed social class, four general attitudes on societal functioning (anomie), and political orientation (right-left). Although self-assessed health is not an attitudinal variable in the classical sense, we believe that the cognitive process required to answer the assessment question is more similar to the cognitive processes of other attitudinal variables that we use than to that for the factual questions that we use because factual questions in our case require simple processing with no extensive recall. The factual questions are employment status, marital status, frequency of church attendance, religious confession, being born in Germany, citizenship, and type of dwelling. We recoded several variables to dichotomous variables. The variable "working for pay" generated from employment status makes a distinction between those who are in paid work and those who are not. "Legal marital status" contrasts legally married persons (irrespective of living together with their spouse) with divorced, never married and widowed persons. Religious confession was recoded in an indicator variable, which takes a value of 1 if a confession was named vs. the value of 0 if no confession was named. The percentage of refusals on this variable is negligible. Citizenship of a specific country was recoded as either having German citizenship or not. A new variable "owner of dwelling" was generated from the variable "type of dwelling."

6.4 Method

We study mode system effects by comparing the estimates from the three surveys. In order to fully understand the processes when comparing the systems of data collection, we first start with a direct comparison of the data from GOPP and the two ALLBUS surveys. In the second step, we compare the GESIS Online Panel Pilot to the subsamples of Internet users from the ALLBUS data in order to eliminate coverage as the possible cause of the differences between GOPP and the two ALLBUS surveys. In the third step, we add weights to compensate for differences in nonresponse between the systems. To ensure that differences between the surveys are not caused by sample composition, we use post-stratification weights based on age, gender and education. The weights were calculated using the inverse proportional fitting (IPF), an iterative weighting strategy that adjusts the marginal distributions of weighted data to the known population margins (Bergmann, 2011).

We treat the distributions of age (five age groups), gender, and education (recoded in three categories) of Internet users in ALLBUS 2010 as reference values (Table E1 in Appendix

E). We would have used population values, however, neither the German Census nor the German Microcensus includes a question on Internet use. Other surveys that could be used as benchmarks (overview in Kaczmirek & Raabe, 2010) as well as the study "Information and communication technologies (ICT) in households" by the German Federal Statistical Office (2011), do not satisfy the quality criteria (e.g., the estimates are based on quota samples). Since the questions were spread through multiple waves in the GOPP, we calculated the post-stratification weights separately for each wave (Tables E2 and E3 in Appendix E). This allowed us to control for attrition and other confounding factors such as experiments with incentives and changing the order of the questionnaires for an experiment on panel conditioning.

Since we use the data from single waves of GOPP without making use of the longitudinal component, no additional panel weights were calculated. The demographic variables in GOPP, which we use for post-stratification here, were collected during the recruitment interview. In rare cases, when the previous multiple interview appointments with a respondent failed or if a respondent was near a break-off, interviewers could ask only about the Internet usage and proceed straight to the recruitment question. In such cases, demographic questions were later asked in the online questionnaires. For the cases with missing data on demographics, we replaced the missing values with information obtained online if it was available. For cases that remained missing, we used single hot deck imputation (Schonlau, 2012) based on gender, that had no missing values. The algorithm identifies all donor observations that have no missing values for the specified variables and replaces all missing values with values from the same randomly chosen donor observation to preserve correlations.

The imputation was performed before forming the age and educational groups. For an educational group of respondents who were still at school, it could not be known what school-leaving degree the respondents would obtain; for those who reported having "other school-leaving degree," it could not be known how their school-leaving degrees relate to the degrees of the German educational system. Therefore, those still in school and those with other school-leaving certificates were set to missing and imputed with the values of the variable "education" before the educational groups were formed. In order to obtain the final weights, post-stratification weights were multiplied with design weights. Design weights in ALLBUS account for an oversampling of persons from East Germany and were provided in both ALLBUS surveys. Design weights for GOPP were calculated using the Gabler-Häder method (Gabler, Häder, Lehnhoff, & Mardian, 2012). According to this method, the design weights equal the inverse probabilities of selection, which take into account the number of telephone lines at which a respondent can be reached and household composition. If the design weight was missing after supplementing it with the data collected online, it was imputed with the modal category from the subsample of respondents willing to take part in the online panel. Design weights were normalized, that is, rescaled to have a mean of 1 and a sum that equals the unweighted number of cases for each questionnaire.

6.5 Results

Table 6.2 presents the estimates obtained from each survey and the results of pairwise *t*-tests for statistical significance of the differences. GOPP differs from ALLBUS surveys in both attitudinal and factual questions. We find statistically significant differences between GOPP and ALLBUS 2010 on 11 out of 12 attitudinal items. The only variable for which no differences were found is the respondents' self-rated financial situation. GOPP differs from ALLBUS 2012 on all attitudinal variables with the exception of 3 items from the anomie-battery, that is, the differences are found on 9 out of 12 attitudinal variables. For factual questions, GOPP differs from the two ALLBUS surveys on all 7 items at a statistically significant level. The differences between the two ALLBUS surveys are found for 10 out of 12 attitudinal variables: current and prospective state of the German economy, current self-rated financial situation, religiosity, the items of the anomie-battery, left-right orientation, and self-assessed social class. For the factual questions, no significant differences between the two ALLBUS surveys were found.

In Table 6.3, we report the results of the comparisons between GOPP and the subsamples of Internet users from the two ALLBUS surveys, excluding the non-Internet-users from the analysis. For several attitudinal variables that showed statistically significant differences between GOPP and the two ALLBUS surveys, no differences are found when GOPP is compared to ALLBUS Internet users only. For Internet users only, the differences between GOPP and ALLBUS 2010 are found on 6 out of 12 attitudinal items: current and prospective state of German economy, health status, religiosity, one item of the anomie-battery, and left-right orientation. The differences from the subsample of Internet users in ALLBUS 2012 are found for 4 attitudinal variables: prospective German economy, current self-rated financial situation, religiosity, and one category of the variable self-assessed social class. We find statistically significant differences between GOPP and the subsample of Internet users in ALLBUS 2010 for all factual variables with the exception of the variable "German citizenship." The differences between GOPP and ALLBUS 2012 Internet users are statistically significant for all 7 factual variables. With the exception of prospective financial situation and self-assessed social class, ALLBUS surveys differ from one another on all attitudinal variables. In sum, when comparing the subsamples of Internet users, ALLBUS surveys differ from each other on 10 out of 12 attitudinal variables and none of the factual variables.

In Table 6.4, we report these comparisons using the weighted data. The set of variables in which GOPP differs from ALLBUS surveys differs from the set of variables that showed significant differences reported in Table 6.3. GOPP differs from either ALLBUS survey in 6 out of 12 attitudinal items. In the following, we refer to the comparisons of GOPP with the subsamples of Internet users in ALLBUS with post-stratification weighting applied to GOPP and ALLBUS 2012.

The variables in which GOPP differs from ALLBUS 2010 are not the same variables in which GOPP differs from ALLBUS 2012. Significant differences between ALLBUS 2010 and GOPP are observed for the following attitudinal variables: state of German economy (both current and prospective), self-rated current financial situation, health status, religiosity, and

Table 6.2. Comparison of the GESIS Online Panel Pilot to two ALLBUS surveys: attitudinal and factual questions.

		GOPP			ALLBUS 2010 full sample			ALLBUS 2012 full sample		
Variable	**Q**	**Estimate**	**SE**	**95% CI**	**Estimate**	**SE**	**95% CI**	**Estimate**	**SE**	**95% CI**
Attitudinal:										
German economy	1	3.400	0.026	[3.350, 3.450]	3.037***	0.016	[3.006, 3.068]	***3.332****	0.015	[3.303, 3.360]
German economy in 1 year	1	2.868	0.029	[2.811, 2.925]	3.014***	0.017	[2.980, 3.047]	***2.662******	0.013	[2.637, 2.688]
Self-rated financial situation	1	3.454	0.027	[3.401, 3.507]	3.411	0.017	[3.379, 3.444]	***3.559*****	0.014	[3.531, 3.587]
Self-rated financial situation in 1 year	1	3.187	0.027	[3.134, 3.241]	3.090**	0.014	[3.064, 3.117]	3.094**	0.011	[3.071, 3.116]
Health status	1	3.679	0.032	[3.616, 3.743]	3.561**	0.021	[3.521, 3.602]	3.562**	0.018	[3.527, 3.597]
Religiosity	4	4.462	0.125	[4.217, 4.707]	5.724***	0.059	[5.608, 5.840]	***6.111******	0.054	[6.006, 6.216]
Self-assessed social class	5									
lower class		2.68%	0.007	[1.66, 4.32]	3.16%	0.003	[2.56, 3.90]	***1.94%***	0.002	[1.53, 2.46]
working class		19.43%	0.017	[16.34, 22.96]	25.75%**	0.008	[24.13, 27.45]	26.21%***	0.008	[24.73, 27.75]
middle class		61.12%	0.020	[57.18, 64.92]	59.74%	0.010	[57.83, 61.61]	61.12%	0.009	[59.41, 62.80]
upper middle class		16.14%	0.014	[13.56, 19.09]	10.84%**	0.006	[9.67, 12.13]	10.08%***	0.006	[9.05, 11.22]
upper class		0.63%	0.003	[0.25, 1.56]	0.51%	0.001	[0.29, 0.89]	0.65%	0.001	[0.42, 1.00]
Life is getting worse	7	78.66%	0.017	[75.20, 81.75]	83.42%**	0.007	[81.93, 84.81]	***77.44%***	0.008	[75.91, 78.90]
Irresponsible to have children	7	35.10%	0.020	[31.36, 39.04]	42.07%**	0.010	[40.19, 43.98]	***35.70%***	0.009	[34.05, 37.39]
Politicians not interested	7	76.86%	0.018	[73.23, 80.14]	80.82%*	0.008	[79.24, 82.31]	***76.18%***	0.008	[74.62, 77.67]
People don't care about others	7	65.30%	0.020	[61.33, 69.07]	73.99%***	0.009	[72.26, 75.64]	***70.91%*****	0.008	[69.27, 72.49]
Left-right orientation	7	4.848	0.070	[4.712, 4.985]	5.218***	0.034	[5.152, 5.284]	***5.045****	0.031	[4.984, 5.106]
Factual:										
Working for pay	2	83.70%	0.014	[80.73, 86.29]	57.80%***	0.010	[55.91, 59.68]	59.76%***	0.009	[58.04, 61.45]
Legal marital status	3	50.47%	0.020	[46.60, 54.34]	57.55%**	0.010	[55.65, 59.42]	57.49%**	0.009	[55.77, 59.20]
Confession	4	64.56%	0.019	[60.74, 68.20]	73.14%***	0.008	[71.49, 74.73]	73.13%***	0.007	[71.65, 74.57]
Frequency of church attendance	4	2.143	0.045	[2.055, 2.231]	2.385***	0.025	[2.335, 2.434]	2.338***	0.023	[2.293, 2.384]
Born in Germany	4	90.72%	0.012	[88.17, 92.76]	84.01%***	0.007	[82.53, 85.39]	85.27%***	0.006	[83.96, 86.49]
German citizenship	4	96.25%	0.007	[94.51, 97.45]	94.07%*	0.005	[93.07, 94.94]	93.93%**	0.004	[93.00, 94.73]
Owner of dwelling	5	47.78%	0.020	[43.86, 51.73]	57.11%***	0.010	[55.20, 58.99]	59.35%***	0.009	[57.64, 61.04]

Note: *p<.05, **p<.01, ***p<.001. Differences between ALLBUS 2010 and 2012 in italic bold, Q short for GOPP questionnaire, SE short for standard error, CI short for confidence interval. N for GOPP (started): Q1=1010; Q2=838; Q3=800; Q4=775; Q5=761; Q7=729. Design weights are applied to all three surveys.

Table 6.3. Comparison of the GESIS Online Panel Pilot to the subsamples of Internet users from two ALLBUS surveys: attitudinal and factual questions.

		GOPP			ALLBUS 2010 Internet users			ALLBUS 2012 Internet users		
Variable	**Q**	**Estimate**	**SE**	**95% CI**	**Estimate**	**SE**	**95% CI**	**Estimate**	**SE**	**95% CI**
Attitudinal:										
German economy	1	3.400	0.026	[3.350, 3.450]	3.086***	0.019	[3.048, 3.124]	***3.355***	0.017	[3.321, 3.389]
German economy in 1 year	1	2.868	0.029	[2.811, 2.925]	3.077***	0.021	[3.036, 3.117]	***2.663**** ***	0.015	[2.633, 2.694]
Self-rated financial situation	1	3.454	0.027	[3.401, 3.507]	3.471	0.020	[3.433, 3.510]	***3.595**** ***	0.016	[3.564, 3.627]
Self-rated financial situation in 1 year	1	3.187	0.027	[3.134, 3.241]	3.182	0.017	[3.148, 3.216]	3.145	0.014	[3.117, 3.173]
Health status	1	3.679	0.032	[3.616, 3.743]	3.823***	0.023	[3.778, 3.868]	***3.725***	0.020	[3.686, 3.764]
Religiosity	4	4.462	0.125	[4.217, 4.707]	5.974***	0.073	[5.832, 6.117]	***6.334**** ***	0.062	[6.213, 6.456]
Self-assessed social class	5									
lower class		2.68%	0.007	[1.65, 4.32]	2.27%	0.003	[1.67, 3.06]	1.58%	0.002	[1.16, 2.14]
working class		19.43%	0.017	[16.34, 22.96]	20.23%	0.010	[18.42, 22.17]	21.81%	0.008	[20.20, 23.52]
middle class		61.12%	0.020	[57.18, 64.92]	63.38%	0.012	[61.08, 65.63]	63.71%	0.010	[61.73, 65.64]
upper middle class		16.14%	0.014	[13.56, 19.09]	13.80%	0.008	[12.22, 15.55]	12.24%*	0.007	[10.93, 13.67]
upper class		0.63%	0.003	[0.25, 1.56]	0.32%	0.001	[0.14, 0.73]	0.66%	0.002	[0.40, 1.10]
Life is getting worse	7	78.66%	0.017	[75.20, 81.75]	82.00%	0.009	[80.10, 83.74]	***75.71%***	0.009	[73.88, 77.45]
Irresponsible to have children	7	35.10%	0.020	[31.36, 39.04]	37.35%	0.012	[35.10, 39.65]	***31.07%***	0.010	[29.22, 32.99]
Politicians not interested	7	76.86%	0.018	[73.23, 80.14]	77.42%	0.010	[75.37, 79.35]	***73.63%***	0.009	[71.75, 75.43]
People don't care about others	7	65.30%	0.020	[61.33, 69.07]	71.72%**	0.011	[69.55, 73.79]	***68.69%***	0.010	[66.74, 70.57]
Left-right orientation	7	4.848	0.070	[4.712, 4.985]	5.143***	0.041	[5.063, 5.222]	***4.937***	0.035	[4.868, 5.006]
Factual:										
Working for pay	2	83.70%	0.014	[80.73, 86.29]	72.41%***	0.011	[70.26, 74.46]	72.11%***	0.009	[70.25, 73.90]
Legal marital status	3	50.47%	0.020	[46.60, 54.34]	56.89%**	0.012	[54.56, 59.19]	57.00%**	0.010	[54.98, 59.00]
Confession	4	64.56%	0.019	[60.74, 68.20]	70.57%**	0.011	[68.47, 72.59]	71.00%**	0.009	[69.20, 72.74]
Frequency of church attendance	4	2.143	0.045	[2.055, 2.231]	2.301**	0.029	[2.244, 2.359]	2.287**	0.026	[2.237, 2.338]
Born in Germany	4	90.72%	0.012	[88.17, 92.76]	87.44%*	0.008	[85.76, 88.94]	86.62%**	0.007	[85.13, 87.99]
German citizenship	4	96.25%	0.007	[94.51, 97.45]	94.53%	0.006	[93.32, 95.52]	93.64%**	0.005	[92.52, 94.59]
Owner of dwelling	5	47.78%	0.020	[43.86, 51.73]	56.64%***	0.012	[54.30, 58.95]	59.19%***	0.010	[57.18, 61.17]

Note: *p<.05, **p<.01, ***p<.001. Differences between ALLBUS 2010 and 2012 in italic bold, Q short for GOPP questionnaire, SE short for standard error, CI short for confidence interval. N for GOPP (started): Q1=1010; Q2=838; Q3=800; Q4=775; Q5=761; Q7=729. Design weights are applied to all three surveys.

Table 6.4. Comparison of the GESIS Online Panel Pilot to the subsamples of Internet users from two ALLBUS surveys: attitudinal and factual questions with post-stratification.

		GOPP			ALLBUS 2010 Internet users			ALLBUS 2012 Internet users		
Variable	**Q**	**Estimate**	**SE**	**95% CI**	**Estimate**	**SE**	**95% CI**	**Estimate**	**SE**	**95% CI**
Attitudinal:										
German economy	1	3.336	0.029	[3.279, 3.393]	3.086***	0.019	[3.048, 3.124]	***3.351***	0.018	[3.317, 3.385]
German economy in 1 year	1	2.844	0.033	[2.779, 2.909]	3.077***	0.021	[3.036, 3.117]	***2.667******	0.016	[2.636, 2.698]
Self-rated financial situation	1	3.394	0.030	[3.334, 3.453]	3.471*	0.020	[3.433, 3.510]	***3.593******	0.017	[3.560, 3.626]
Self-rated financial situation in 1 year	1	3.168	0.031	[3.107, 3.228]	3.182	0.017	[3.148, 3.216]	3.157	0.015	[3.128, 3.186]
Health status	1	3.606	0.038	[3.531, 3.681]	3.823***	0.023	[3.778, 3.868]	***3.748*****	0.020	[3.709, 3.787]
Religiosity	4	4.585	0.145	[4.300, 4.870]	5.974***	0.073	[5.832, 6.117]	***6.341******	0.063	[6.218, 6.466]
Self-assessed social class	5									
lower class		3.36%	0.010	[1.85, 6.03]	2.27%	0.003	[1.67, 3.06]	1.55%	0.003	[1.13, 2.13]
working class		24.81%	0.023	[20.52, 29.65]	20.23%	0.010	[18.42, 22.17]	22.21%	0.009	[20.55, 23.97]
middle class		59.36%	0.024	[54.60, 63.94]	63.38%	0.012	[61.08, 65.63]	63.24%	0.010	[61.21, 65.22]
upper middle class		12.02%	0.012	[9.87, 14.56]	13.80%	0.008	[12.22, 15.55]	12.36%	0.007	[11.03, 13.83]
upper class		0.45%	0.002	[0.18, 1.16]	0.32%	0.001	[0.14, 0.73]	0.64%	0.002	[0.38, 1.09]
Life is getting worse	7	79.51%	0.019	[75.47, 83.04]	82.00%	0.009	[80.10, 83.74]	***75.58%***	0.009	[73.70, 77.36]
Irresponsible to have children	7	40.41%	0.024	[35.87, 45.13]	37.35%	0.012	[35.10, 39.65]	***31.06%******	0.010	[29.17, 33.01]
Politicians not interested	7	79.28%	0.019	[75.38, 82.70]	77.42%	0.010	[75.37, 79.35]	***73.56%*****	0.010	[71.64, 75.40]
People don't care about others	7	68.45%	0.022	[64.09, 72.50]	71.72%	0.011	[69.55, 73.79]	***68.63%***	0.010	[66.64, 70.55]
Left-right orientation	7	4.945	0.080	[4.788, 5.102]	5.143*	0.041	[5.063, 5.222]	***4.946***	0.036	[4.876, 5.016]
Factual:										
Working for pay	2	83.55%	0.016	[80.28, 86.37]	72.41%***	0.011	[70.26, 74.46]	73.60%***	0.009	[71.75, 75.37]
Legal marital status	3	49.87%	0.023	[45.30, 54.44]	56.89%**	0.012	[54.56, 59.19]	56.23%*	0.011	[54.16, 58.28]
Confession	4	67.29%	0.021	[63.10, 71.21]	70.57%	0.011	[68.47, 72.59]	71.29%	0.009	[69.46, 73.04]
Frequency of church attendance	4	2.175	0.052	[2.073, 2.277]	2.301*	0.029	[2.244, 2.359]	2.286	0.026	[2.234, 2.337]
Born in Germany	4	91.04%	0.013	[88.08, 93.32]	87.44%*	0.008	[85.76, 88.94]	86.25%**	0.008	[84.69, 87.67]
German citizenship	4	96.50%	0.008	[94.63, 97.73]	94.53%*	0.006	[93.32, 95.52]	93.50%**	0.006	[92.33, 94.50]
Owner of dwelling	5	45.38%	0.023	[40.86, 49.98]	56.64%***	0.012	[54.30, 58.95]	58.06%***	0.010	[56.99, 60.10]

Note: *p<.05, **p<.01, ***p<.001. Differences between ALLBUS 2010 and 2012 in italic bold, Q short for GOPP questionnaire, SE short for standard error, CI short for confidence interval. N for GOPP (started): Q1=1010; Q2=838; Q3=800; Q4=775; Q5=761; Q7=729. Weighted data: ALLBUS 2010 design, GOPP & ALLBUS 2012 design and post-stratification.

6

left-right orientation. The set of attitudinal variables with statistically significant differences between GOPP and ALLBUS 2012 excludes the current state of German economy and left-right orientation but includes two of the items of the anomie-battery. However, the two ALLBUS surveys show significant differences for the same 10 out of 12 items reported in Table 6.3.

For factual questions, GOPP differs from ALLBUS 2010 on 6 out of 7 items: on all factual questions with the exception of confession. Estimates from GOPP differ from ALLBUS 2012 on 5 out of 7 items, the exceptions being confession and frequency of going to church. As is the case with unweighted data, there are no significant differences between the two ALLBUS surveys for any of the factual questions. This finding indicates the presence of the mode system effect in our data. The fact that the two ALLBUS surveys differ on attitudinal questions shows that attitudes, being unstable constructs, would not have allowed us to single out the mode system effect. We see that two surveys with identical recruitment and design features do not differ in factual questions, whereas the difference between the online and the face-to-face surveys is obvious.

However, finding statistically significant differences between the reference surveys is not the only indicator of data quality. If we use only the reports of significance, the overall differences reported in Table 6.4 between GOPP and ALLBUS 2010 (12 out of 19 items), between GOPP and ALLBUS 2012 (11 out of 19), and between the ALLBUS surveys (10 out of 19), seem similar. However, the count of statistically significant differences depends on the number of variables we choose to study and does not provide an insight into the magnitude of the differences. In addition to statistical significance, Biemer (1988) recommends examining the effect size, the direction of the difference, and the violations of the underlying assumptions for the mode comparison study, which could explain the magnitude of the difference. In Table 6.5, we report the unweighted and weighted overall and absolute mean effect sizes. The effect sizes standardize the comparisons between the means and proportions reported in Table 6.4 and allow for a better estimation of error because they not only take into account the difference (bias) but also the sample sizes and the precision of estimates (variance).

From Table 6.5, we can conclude that the directions of all mean effect sizes are the same. For weighted mean effect sizes over all questions the difference between GOPP and ALLBUS 2010 (-0.063) and the difference between GOPP and ALLBUS 2012 (-0.036) are larger than the difference between the two ALLBUS surveys (-0.030). However, the difference between GOPP and ALLBUS 2012 has almost the same magnitude as the difference between the two face-to-face surveys. If we look at the effect sizes split by the type of the questions, the results mimic what we have seen when comparing the frequency counts of statistically significant results. For factual questions, the difference between the two face-to-face surveys is minimal (0.001) and almost identical for online vs. face-to-face (0.048 and 0.050). For attitudinal questions ALLBUS 2012 is closer to GOPP (-0.031) than to ALLBUS 2010 (-0.068). The difference between two face-to-face surveys on attitudinal questions is larger than the difference between GOPP and ALLBUS 2012 (-0.042 vs. -0.031). Although effect sizes differ between the surveys and between the question types, they are very small, the

Table 6.5. Mean effect sizes and inverse variance weights across the surveys.

	GOPP vs. ALLBUS 2010 Internet users		GOPP vs. ALLBUS 2012 Internet users		ALLBUS 2012 vs. 2010 Internet users	
Variable	**d**	**w**	**d**	**w**	**d**	**w**
Attitudinal:						
German economy	0.295	625.000	-0.017	688.705	0.307	1051.525
German economy in 1 year	-0.247	623.053	0.206	682.128	-0.488	1019.368
Self-rated financial situation	-0.087	630.120	-0.229	684.932	0.143	1062.699
Self-rated financial situation in 1 year	-0.017	626.566	0.014	684.463	-0.034	1052.632
Health status	-0.206	621.504	-0.136	679.348	-0.075	1064.963
Religiosity	-0.407	530.786	-0.523	568.182	0.117	1059.322
Self-assessed social class						
lower class	0.222	49.998	0.437	48.603	-0.214	64.045
working class	0.145	311.915	0.080	346.380	0.065	571.755
middle class	-0.094	416.840	-0.090	452.284	-0.003	800.000
upper middle class	-0.087	191.022	-0.018	201.086	-0.070	393.236
upper class	0.189	7.004	-0.195	9.100	0.384	13.942
Life is getting worse	-0.088	269.469	0.125	305.344	-0.213	552.181
Irresponsible to have children	0.071	406.669	0.225	428.266	-0.154	776.398
Politicians not interested	0.060	284.738	0.175	310.945	-0.115	626.174
People don't care about others	-0.086	362.450	-0.005	394.477	-0.082	719.424
Left-right orientation	-0.107	509.684	-0.001	549.451	-0.113	1022.495
Factual:						
Working for pay	0.364	280.741	0.331	296.121	0.033	692.042
Legal marital status	-0.156	453.309	-0.141	490.196	-0.015	861.326
Confession	-0.085	383.877	-0.104	413.736	0.019	720.461
Frequency of church attendance	-0.097	539.374	-0.084	582.072	-0.012	1062.699
Born in Germany	0.209	157.109	0.266	169.033	-0.058	397.931
German citizenship	0.258	66.934	0.359	72.417	-0.101	193.836
Owner of dwelling	-0.250	435.730	-0.282	468.604	0.032	856.164
Weighted mean effect size:						
attitudinal questions		-0.068 (0.012)		-0.031 (0.012)		-0.042 (0.009)
factual questions		-0.048 (0.021)		-0.050 (0.020)		-0.001 (0.014)
overall		-0.063 (0.011)		-0.036 (0.010)		-0.030 (0.008)
Weighted mean absolute effect size:						
attitudinal questions		0.157 (0.012)		0.141 (0.012)		0.150 (0.009)
factual questions		0.180 (0.021)		0.185 (0.020)		0.028 (0.014)
overall		0.163 (0.011)		0.153 (0.010)		0.115 (0.008)

Note: *d* is the unweighted effect size, *w* is the inverse variance weight, standard errors in parentheses. The formulas are in Appendix E.

cut-off point for *d* being 0.1 (Ferguson, 2009, p. 533). The absolute weighted mean effect sizes reported in the end of Table 6.5 do not account for the direction of the effect but only for the magnitude. Here we again see that for attitudinal items GOPP estimates are closer to ALLBUS 2012 estimates, whereas for factual questions the difference between the two face-to-face surveys is much smaller than between either ALLBUS survey and GOPP, and the overall difference between GOPP and either ALLBUS survey is larger than between the ALLBUS surveys. Although absolute mean effect sizes are larger than mean effect sizes, they are still small (they do not exceed 0.2). From a practical point of view, the mode system effects between the surveys are therefore negligible when judged by their effect sizes and small when judged by the absolute effect sizes.

6.6 Discussion

In this chapter, we study the mode system effect comparing the data from a probability-based telephone-recruited online panel with the subsamples of Internet users from two face-to-face surveys. All three sources implemented questions with identical wording and we control for differences in sample composition to single out the mode system effect. We distinguish between factual and attitudinal questions. We hypothesize that the effects are more pronounced for attitudinal questions than for factual questions. Our hypothesis finds no support. There are differences between the online collected data and interviewer collected data on both types of questions. However, our distinction is important. We find that for factual questions both face-to-face surveys differ from the online panel but do not differ significantly from each other. This conclusion is supported by the analysis of the effect sizes. For attitude questions, the difference between the interviewer-administered surveys is larger than that of a face-to-face survey (ALLBUS 2012) and the online panel. We attribute this result to the instability of attitudes. However, alternative explanations are possible. For example, the reference period of ALLBUS 2012 is closer to the online panel administered in 2011-2012 than to ALLBUS 2010. The effect sizes are relatively large for the questions that refer to the economic situation. The differences between the surveys on those variables might be attributed not to the mode system effect but to the real change given the financial upheavals between 2010 and 2012 in Germany. To check the robustness of our results, we recalculated the effect sizes when the four variables that refer to the economic condition are excluded. This did not affect our overall conclusions, but it provides us with more accurate measures of the mode system effect.[32]

Our findings have important implications for estimating mode system effects. First, we have shown the importance of having more than one reference survey. Both face-to-face surveys serve in our case as control surveys for the online panel and allow us to make inferences on whether differences could be attributed to the mode or not. We use two

[32] Weighted mean effect sizes without the economic items: overall: -0.082 (0.013), -0.048 (0.012), -0.036 (0.009); absolute: 0.164 (0.013), 0.168 (0.012), 0.073 (0.009); for attitudinal items: -0.102 (0.016), -0.047 (0.015), -0.058 (0.011); absolute for attitudinal items: 0.155 (0.016), 0.157 (0.015), 0.101 (0.011).

surveys with equal recruitment procedures. An alternative would be to use two waves of the panel survey. However, we believe that having two unrelated cross-sections in this case is a better trade off than panel data which have specific problems of attrition and panel conditioning. Second, we show the importance of distinguishing between types of questions. The differences between online and interviewer-administered modes are well-documented for sensitive questions (e.g., meta-analysis by Tourangeau, Conrad, & Couper, 2013, p. 142).[33] We draw on another dimension, in which a distinction is made based on the cognitive demands that questions place on a respondent. Our findings are in line with Klausch, Hox, and Schouten (2014), who also report that mode effects depend on the question type. Third, we find a common denominator for the comparison of mode system effects by reporting the effect sizes. The magnitude of the mode system effect when judged by the effect sizes is small. However, researchers who use the data from large surveys might be misled if in they rely solely on significance testing. We encourage other researchers to report effect sizes. This would allow for comparing our results to similar studies based on different data. Finally, for analysts of data, the conclusion to draw from our analysis is that mode system effects differ across question types. Thus, data users should have different concerns about mode effects when analyzing only attitudinal, only factual, or both types of items.

[33] We did not have sensitive items to use for our analysis. However, for religiosity, which was measured on a 10-point scale and the mean values of which we report in Table 6.3, we find that the percentage of online panel respondents who choose the answer “not religious” is much higher than for interviewer-administered surveys (29.44% with SE=0.021, CI=[25.56, 33.64] for GOPP and 7.72% with SE=0.007, CI=[6.54, 9.10] for ALLBUS 2010, 6.27% with SE=0.005,CI=[5.32, 7.38], p<.001 for GOPP vs. either ALLBUS). This could be indicative of social desirability.

Chapter 7

Conclusion

There are many types of surveys, and surveys serve a variety of purposes. The goals of academic and governmental surveys are to understand human behavior and social reality at large, among other things, to be able to make well-informed policy decisions. In order to use information from surveys for these purposes, data users need to be sure that this information is of high quality.

Conducting high-quality surveys is expensive. The dominant mode of administration for high-quality surveys is the administration by an interviewer, preferably, in person (face-to-face probability-based surveys). Online modes of data collection, especially online panels, offer cost reductions. Probability-based online panels are generally preferred by the scientific community. If a design is applied in which the first wave of a probability-based panel is carried out with an interviewer (by telephone or face-to-face) and the subsequent data collection is performed in the web mode, the costs are only high for the one-time recruitment. In the case of mail recruitment, the costs are lower. However, the question remains: Is the quality of data in probability-based online panels "good enough," that is, comparable to face-to-face or telephone surveys?

Some design peculiarities of probability-based online panels may threaten the correctness of the conclusions based on the data they generate. The first question that the researcher who wants to recruit a probability-based panel faces is whom to include (sampling). There is no complete list of email addresses of the general population available; therefore, the researcher needs to rely on alternative sources, such as population registers, samples of telephone numbers (Random Digit Dial), or address-based sampling. The second question is how to reach non-Internet users (coverage). If the goal of the study is to reach conclusions about the entire population and Internet users differ from non-users in important aspects, the results of the study will be misleading. In that case, an alternative completion mode or equipment with Internet access needs to be offered to those who do not have access to the Internet.

When these two issues are resolved, the selected individuals may still refuse to participate for various reasons (nonresponse). They may not be able to participate in an online survey either because their computer skills are not sufficient or because they are not comfortable completing surveys on the Internet since this usually requires some additional actions, such as clicking on the link in the email invitation or typing the link into the browser, registering with a portal, etc. If potential respondents are able to participate, this does not mean they will be willing to do so. Finally, if they are able and willing to participate, there is still a possibility that they will not start the online survey. The respondents have to actively start the survey in the mode different from the one in which they were contacted and at a later time after they were first reached, and what happens during the agreement and the actual participation is usually unknown to the researcher. Furthermore, completing the first online survey does not mean that respondents will continue further participation in a panel. Panel members may skip one or several waves or quit the participation entirely (attrition). Additionally, there are important measurement issues. For example, the completion of multiple online surveys can lead to respondents' learning to recognize "shortcuts" in the questionnaire (panel conditioning). All of the above will negatively affect the data collected in a probability-based online panel.

The central question of this dissertation is: How good is the data from probability-based online panels? The five studies in this dissertation focus on the quality of the final estimates produced by an online panel and provide insight into whether and how the processes of nonresponse, attrition, and panel conditioning influence data quality. All studies in this dissertation use the data generated by one offline-recruited probability-based online panel of Internet users in Germany: the GESIS Online Panel Pilot. Each study constitutes one chapter and highlights a different aspect of data quality.

7.1 Summary of the results

Chapter 1 provided a description of the concepts of data quality, total survey error, and the mode system approach, which are central for the studies in this dissertation. Moreover, Chapter 1 showed how the error types relate to the recruitment process and the subsequent data collection processes of a probability-based online panel. In Chapter 2, the telephone recruitment and the online interviewing of the GESIS Online Panel Pilot were described in detail. The central question of Chapter 2 was whether the data collected within an online panel was comparable to "gold standard" surveys. Two high-quality face-to-face administered surveys acted as benchmarks: the German General Social Survey ("ALLBUS") and the German sample of the European Social Survey (ESS). We compared subsamples of Internet users from these two surveys to the data that were collected in the GESIS Online Panel Pilot. The results of the comparison of point estimates showed differences on almost all of the variables. However, most of these differences were within a few percentage points. Users of survey data may not be as interested in descriptive statistics as they are in modeling social processes using these data. Thus, in order to assess the quality of data ("fitness for use"), providing comparisons of point estimates does not suffice. We used data from all three surveys for multivariate analysis and compared the regression coefficients. Although the magnitude of these coefficients differed, the direction was the same for most of the coefficients. It should be noted that for the point estimates and the results of multivariate analysis, we also observed significant differences between the two reference surveys. Finally, this chapter demonstrated that the strategy of reducing undercoverage and nonresponse by using the post-survey adjustment (post-stratification weighting) did not reduce the differences between the online panel estimates and the estimates from the reference surveys. Overall, the result of Chapter 2 is that the data collected in the online panel is comparable to the benchmark surveys. While Chapter 2 concentrated on the "final products" of the survey (i.e., point estimates), three subsequent chapters investigated the processes that generate these products.

Chapter 3 concentrated on the selectivity effects that occurred during both the telephone recruitment and the actual participation in the online panel. The recruitment process for probability-based panels has multiple stages, at each of which the decision can be made not to participate. If persons that are selected for the panel but who do not participate provided answers that differ from those provided by persons who participated, the conclusions drawn from the data may be misleading. The questions that Chapter 3 addressed were: Do persons who agree to participate in the panel during the recruitment interview differ from those who refuse? Do participants differ from those who fail to participate? If so, what are the differences? The results showed that the willingness to participate and the actual online participation were associated with different sets of demographic characteristics. Online survey experience was predictive of online participation and especially of willingness to participate. Once online survey experience and Internet-use-related characteristics were controlled for, differences on some demographic characteristics between individuals willing and individuals nonwilling to participate and between respondents and nonrespondents no longer held. Among the attitudinal variables, the level of trust was predictive of the willingness to participate but not of the participation itself. High incentives and being recruited by the in-house telephone lab also predicted the agreement to participate in the panel. According to Groves, Singer, and Corning (2000), the decision to participate in a survey depends on the leverage (importance) that respondents assign to various aspects of the survey request and how salient a particular attribute is. We hypothesized that the influence of the fieldwork agency would diminish at the stage of actual online participation, because the contact with an interviewer was not present (salient) at this stage. However, the effect of the fieldwork agency also persisted at the stage of actual participation.

Chapter 4 and Chapter 5 addressed problems that are specific to panel surveys. Chapter 4 focused on panel attrition, which was defined as a failure to complete any number of panel questionnaires after having completed at least one of them. Continuous participation of respondents is essential for the quality of a panel survey. The representativeness of a panel is threatened if those who do not participate are different from those who do, because the chances that the determinants of attrition are related to the outcomes of interest are high for multi-topic panels. Furthermore, after several waves, the number of cases can become too small, which affects the variance of estimates and thereby decreases the inferential capabilities of a panel. In order to prevent respondent attrition, it is important to understand what causes panel attrition. In Chapter 4, two approaches of respondents' motivation to continue participation or to fail to participate were contrasted: experience during the previous survey and monetary incentives. Using the time-varying measures of survey experience that were collected in every wave, we found that the likelihood of participation was lessened for respondents who rated their survey experience as burdensome. On the other hand, positive survey experience did not guarantee participation in the following wave. However, respondents who rated the previous survey as important for science were less likely to attrite. Monetary incentives generally reduced the likelihood of attrition: higher incentive amounts were associated with lower attrition. However, respondents needed to actively initiate payment (in the form of cash, vouchers, or charity donation), and once we controlled for payment, the incentive amount was not decisive. Nevertheless, participants in higher incentives groups were also more likely to initiate payment. We found no evidence that incentives compensated for a negative survey experience; instead they served as an additional bonus to a pleasant survey experience. We also found that respondents with burdensome experience were more likely to attrite monotonically, whereas the pattern of irregular participation was not associated with negative panel experience. In line with the previous findings from the literature, we found that attrition was weakly related to respondents' demographic characteristics.

Chapter 5 investigated panel conditioning, that is, the possibility that survey responses are affected by participation in the previous waves of a panel. Panel conditioning may occur as a result of respondents' learning the questionnaire content or the respondents' learning the survey procedure. In this chapter, we concentrated on the latter. Learning the survey procedure can increase or decrease data quality. As a result of previous interviewing, respondents may provide more honest, less socially desirable answers (advantageous conditioning). They may also learn how to manipulate the survey instrument: for example, to recognize filter questions or foresee repeated elements of the questionnaire and respond accordingly in order to minimize the burden of answering follow-up questions (disadvantageous conditioning). Using two field experiments that employed rotating the questionnaire content, we investigated both advantageous and disadvantageous conditioning. Limited advantageous and no disadvantageous conditioning were found.

Chapter 6, much like Chapter 2, focused on final survey estimates. In this final chapter, we also assessed the quality of data from the online panel by comparing estimates from the online panel to estimates from two face-to-face reference surveys. However, this chapter differed in some important aspects from Chapter 2. Here, a probability-based offline-recruited online panel was considered to be a specific type of the system of data collection, that is, "an entire data collection process designed around a specific mode" (Biemer & Lyberg, 2003, p. 208). The estimates are the final product of the set of specific design decisions and practices, which include but are not limited to the mode of data collection. Chapter 6 estimated these mode system effects (in contrast to pure mode effects) and sought to answer the question "What would happen if a researcher decided to change the data collection mode from interviewer administration to online surveys?" We used two ALLBUS surveys as reference surveys. These surveys were two cross-sectional surveys in which data collection took place in 2010 and 2012. Both surveys had almost identical recruitment protocols; therefore, they constituted a system of data collection. The GESIS Online Panel Pilot was the other system of data collection. We used questions with identical wordings that were present in both ALLBUS surveys and that were asked in the online questionnaires of the panel. The differences in sample composition were accounted for by weighting. The findings indicated that all three surveys differed from each other in the attitudinal measures. For the factual questions, differences were found between the online panel and the reference surveys. However, no differences were found between the two reference surveys for factual questions. This is a clear indication of a mode system effect. Another important finding of this final chapter was the small magnitude of the mode system effects. Since the estimates from the panel accumulated all potential errors (nonresponse, attrition, mode change between the recruitment and the panel waves), the small magnitude of the difference is further evidence that the data from an online panel is fairly comparable to the surveys with traditional modes of data collection.

7.2 Limitations and suggestions for future research

The five studies in this dissertation have some inherent limitations that will be discussed in this section together with some suggestions for future research.

This dissertation focused on the quality of survey data. In this respect, it is important to note that survey quality is not an absolute; rather, it is relative to user requirements and producer constraints, such as schedule or survey budget (Biemer, 2010; Biemer & Lyberg, 2003). The method of quality assessment that Chapter 2 and Chapter 6 undertook also viewed the data quality relative to the two reference surveys. This presumed that the data obtained by the reference surveys was of high quality (accurate). The assumption that the data from face-to-face surveys is error-free is a strong one, and future investigations that evaluate the reference surveys would be useful. A related point is that the choice of measures that were used for comparisons with the two benchmark surveys was guided mainly by their availability (i.e., comparable or identical variables had to be present in ALLBUS and ESS or in the two ALLBUS surveys). We chose variables that we considered not to evoke socially desirable responses since in previous studies substantial differences were found between face-to-face and online mode in this respect (see, for example, meta-analysis in Tourangeau, Conrad, & Couper, 2013, pp. 132-133). However, the analyses presented in Chapters 2 and 6 would have benefited from more comparisons of model estimation. Ideally, several constructs should be measured in all of the surveys to analyze how robust the results are. The solution would be to employ large-scale experiments (e.g., recently reviewed by Callegaro, Villar, Yeager, & Krosnick, 2014), that would carefully consider which measures should be included. If the research community could agree on a set of constructs measured by a set of variables that should be collected in surveys for comparison purposes, this would allow for a better comparability of the studies investigating online panel data quality. However, such a consensus seems unlikely in the near future.

Due to the pilot nature of the project in which the data were collected, several constraints were imposed, such as budget constraints that influenced design decisions in recruitment and online interviewing. As a result, some aspects could not be covered in this dissertation. As opposed to sampling from registers, telephone recruitment did not allow for the auxiliary frame information to be used for studying nonresponse. Thus, in Chapter 3, nonresponse to the recruitment interview could not be investigated. Furthermore, the design decision not to interview the non-Internet population prevented the analysis of coverage error as well as the interplay of these two types of errors. It is not clear how the initial nonresponse and nonresponse at subsequent recruitment errors interact. The representativeness of the panel, nevertheless, need not be worsened by initial nonresponse if various stages of recruitment produce countervailing biases. It must be noted, however, that having a population register and being able to use this register for sampling purposes as well as the availability of frame information are more of an exception than a rule in the survey world. Moreover, despite the low response rate (see Chapter 2 for details), the data from the online panel proved to be fairly comparable to that of face-to-face surveys with higher response rates typical for Germany at the moment (see Chapters 2 and 6).

The analysis in Chapter 4 might be affected by the rather small number of cases and the short time period of eight months. The small number of attriting cases did not allow us to use the full potential of the panel analysis methods. A longer period of observation could have led to different results concerning the distribution participation patterns (cf. Lugtig, 2012, 2014). Nonetheless, Chapter 4 provided additional evidence for the longitudinal effects of incentives. The literature on the long-term effects of incentives does not provide straightforward conclusions. Incentives have been shown to cement respondents' loyalty (e.g., Jäckle & Lynn, 2008; Laurie, 2007; Laurie & Lynn, 2009), however, ceasing to provide or reducing the incentive did not lead to decreases in participation (e.g., Lengacher, Sullivan, Couper, & Groves, 1995; Singer, Van Hoewyk, Gebler, Raghunathan, & McGonagle, 1999). The simultaneous study of incentives and survey experience in this chapter may shed some light on this nexus, which should be investigated in future studies. Chapter 4 also provided an additional case of analyzing time-varying measures of which the attrition literature currently falls short. The limitations discussed above call for replications of the attrition study presented in this dissertation. It would furthermore be useful to replicate this study in interviewer-administered panel surveys. The influence of an interviewer, which is important in face-to-face panels (cf. Behr, Bellgardt, & Rendtel, 2005; Hill & Willis, 2001; Laurie, Smith, & Scott, 1999; Nicoletti & Peracchi, 2005; Watson & Wooden, 2009; Zabel, 1998), might lead to different results.

Chapter 5 would have benefitted from the analysis of panel conditioning due to the learning of the questionnaire content. An ideal study would investigate the effects of both learning the content and learning the survey procedure. In order to separate these two processes using the design of the experiments in Chapter 5 (i.e., random assignment and changing the order of the questionnaires), two different sets of measures need to be used. One set of measures needs to be presented to both groups once to study panel conditioning due to learning the survey procedure and another set of measures needs to be presented multiple times to the experimental group and once to the control group. Less important, but also worth mentioning is that the small number of cases did not allow for subgroup analysis and negatively affected the power of the study. However, the analysis of paradata for robustness checks might have offered a solution to this problem. To this end, this chapter provided a valuable contribution to survey methodology and should encourage the use of paradata in conditioning studies.

The overarching theoretical framework for the studies in this dissertation was the framework of survey quality (Biemer & Lyberg, 2003) and two concepts related to it: total survey error (e.g., Groves & Lyberg, 2010) and whole systems approach (Biemer, 1988). These approaches served well in Chapters 2 and 6. However, Chapters 3, 4, and 5 mainly relied on additional theoretical approaches: the framework of survey cooperation in household surveys (Groves & Couper, 1998), the leverage-saliency theory (Groves et al., 2000), the benefit-cost theory of survey participation (Singer, 2011), and satisficing (Krosnick, 1991). This is plausible because Chapters 2 and 6 focused on data quality, whereas the other chapters focused on the processes that generate the data. Among other commonalities, the frameworks used for Chapters 3–5 view the respondent as a rational actor "weighting" survey benefits against costs in his or her decision to participate or to choose "shortcuts" in answering the questions (satisficing). Although studies in Chapters 3–5 are placed in the context of total survey error (focus either on nonresponse or on measurement), the errors in these chapters were not quantified, and no attempt was undertaken to link the error types. It was not the goal of this dissertation to provide extensive empirical tests of the underlying theoretical mechanisms. This is a direction for future work and would fall under what Groves and Lyberg suggested as one possible agenda for the future of total survey error, that is, the "specification of survey error models arising from social and behavioral science findings in survey behavior" (Groves & Lyberg, 2010, p. 871).

Finally, this dissertation considered it established knowledge that the costs of telephone recruitment and online interviewing are generally lower than the costs of face-to-face recruitment and interviewing. From the point of view of survey practice, a careful analysis of cost savings in data collection and processing would have made a stronger argument for the total quality of online panel survey data.

7.3 Some thoughts about the future of probability-based online panels

In this dissertation, online interviewing has often been contrasted with the "traditional" modes of data collection, a term that here refers to face-to-face and telephone surveys, but naturally, mail surveys that were not the focus of this particular work also fall under a category of a traditional data collection mode. In some countries, probability-based online panels have established themselves as a cost-effective alternative to face-to-face and telephone surveys of the general or Internet-using population. In other countries, they are beginning to emerge or are under consideration (Nicolaas, Calderwood, Lynn, & Roberts, 2014).

While probability-based online panels have not yet become a traditional (i.e., widely accepted) mode, at the time of writing this mode ironically seems to be "old" for the survey world, which is influenced by accelerating technological advances. The predictions of the early 2000s that online interviewing would replace face-to-face and telephone data collection (which were addressed by Couper (2000) with much skepticism), are currently being replaced by the concerns that online interviewing, together with other surveys, will be replaced by new developments, such as "big data" (Couper, 2013). The limits of using big data for survey research make it unlikely that it will replace sample surveys (Couper, 2013). The replacement of high-quality surveys by other "surveys without questions," (i.e., register data), is also unlikely (Bethlehem, 2008). There is reason to believe that probability-based online panels will continue to exist as stand-alone surveys or as a part of a mixed-mode design (more on the latter in de Leeuw & Hox, 2011). In fact, probability-based online panels are not that new: the Dutch Telepanel that was based on a probability sample and provided respondents willing to participate with equipment for Internet interviewing was set up almost three decades ago (Saris, 1998). However, the systematic investigations of data quality in online panels are relatively new (see the recent review in Callegaro et al., 2014). Efforts to evaluate the data quality should continue.

It is important to apply the acquired knowledge about the quality of data in probability-based online panels when building new panels or integrating this method of data collection into an existing panel that uses another (traditional) mode. The results of the previous studies can guide the design decisions in prospective probability-based online panels. At least three scenarios are possible in which probability-based panels can be implemented and can benefit from the studies on data quality.

Online panels for the general population: Although the question concerning online panels of the general population in some research settings is still “Is it feasible?”, it is generally turning into the question “How?” (for example, the recent 2013 Conference of the European Survey Research Association contained three sessions about online panels for the general population).[34] Some argue that future increases in Internet access rates will resolve the issues of coverage, and as more people use the Internet, response to an online survey (or multiple surveys in the case of a panel) will also increase. This optimism seems preliminary at least for the near future. For example, in Finland, where the current Internet penetration is 89%, the recruitment rate for an online panel with a sample drawn from a population registry was still in the single digits (Grönlund & Strandberg, 2014). To be fair, the recruitment method might also have played a role: in this case, it was completed by mail. For the LISS Panel, which was also built in a country with high Internet penetration (the Netherlands) and whose recruitment was interviewer-administered, the proportion of persons willing to participate was 63% (Scherpenzeel & Das, 2011).

Furthermore, having Internet access does not automatically mean being familiar with navigating web pages or being comfortable with completing online surveys. For example, in the GESIS Panel, which is a successor of the GESIS Online Panel Pilot, about 22% of Internet users asked to be sent a mail questionnaire during a face-to-face recruitment interview. Overall, the proportion of infrequent Internet users who are automatically assigned to the mail mode and Internet users explicitly asking to participate by mail, is as high as 25% among the Internet users who are willing to participate in the GESIS Panel.[35]

Thus, probability-based online panels aimed at generalizing the results to the entire population will have to continue to provide equipment and Internet access to non-users or offer an alternative mode. Referring to the latter example (the GESIS Panel), those considering building probability-based online panels should look more into the option of offering an alternative mode with which respondents feel most comfortable. This strategy, however, imposes some constraints on measurement: the use of interactive features, pictures, colors, the feeding forward the information from previous interviews (dependent interviewing), randomization of questions, extensive filters, and the numerous experimental conditions that can all be used in the online mode would be difficult or impossible to implement in the offline mode. Moreover, providing similar stimuli to avoid the effects on measurement (Dillman, Smyth, & Christian, 2009) might be challenging.

---

34 http://www.europeansurveyresearch.org/conf/ESRA_novo.pdf.

35 Own calculations. *GESIS (2014): GESIS Panel - full version. GESIS Data Archive, Cologne. ZA5664 Data file Version 1.0.0, doi:10.4232/1.11877* .

Providing equipment to non-Internet users or offering an alternative mode (mixed-mode design) increases costs: Even for a postal mail mode that requires lower costs than the interviewer-administered modes, additional printing, postage, and processing costs can be substantial. Thus, for studies in which generalizations to the full population are not crucial, *probability-based panels of Internet users* may be a viable option that allows more flexibility in measurement, complex filtering, and other benefits described above without the added costs caused by offline respondents. The data from such panels would satisfy the needs of the scientific community for valid inference, which can only be obtained from nonprobability online panels in a limited number of cases (cf. Callegaro et al., 2014). Finally, there is also a scenario of *switching from a traditional mode to an online mode* completely or for certain groups of respondents. In such instances, it is crucial to know whether changes in trends are actual changes in the population or are due to the change of a survey mode. In other words, data users need to be sure that the data are coherent, that is, that the estimates from different sources can be combined (Biemer, 2010). To estimate the mode effect, experimental studies need to be administered with a random assignment to the mode (de Leeuw & Hox, 2011). It is likely that the mode will not be the only factor that varies, so the studies investigating mode system effects will be needed as well. In sum, it will be fruitful for the future research agenda for probability-based panels to incorporate mode effects studies. There are numerous combinations of mixing the modes (de Leeuw, 2005; de Leeuw & Hox, 2011; Dillman et al., 2009), so there are a lot of future research opportunities for probability-based panels in this respect.

A further point that needs to be addressed is that one mode (online surveys) does not guarantee the completion on a similar device by all respondents. The increased use of handheld devices such as smartphones and tablets poses new questions about the data quality in online surveys completed on mobile devices (Callegaro, 2013a; Dillman et al., 2009). This area is gaining increased attention (see, for example, the programs and proceedings of the AAPOR, ESRA, and the General Online Research (GOR) Conference from recent years) and will continue to be investigated in the future since the effects of survey taking via mobile devices on measurement are yet not well understood. Couper (2013) suggests that the challenges that mobile data collection brings be addressed by developing better questionnaires and better measures to reduce the survey length or burden because lengthy questionnaires, which are typical for the face-to-face mode, are not feasible when completing surveys on mobile devices or in online panels. Additionally, it is foreseeable that more attention will be directed at securing continuous participation. Since online panels employ short questionnaires, it is imaginable that certain constructs that require extensive measures would be spread over several questionnaires because these constructs are too long or too burdensome to complete in one survey. Studies aimed at investigating what actions of the panel management foster respondents' engagement needed for continuous participation, such as the recent study by Scherpenzeel and Toepoel (2014), are needed for understanding attrition and will complement the studies of attrition that mostly assess the bias that attrition creates. Another solution to the problem of lengthy questionnaires would be to improve the validity and reliability of single-item measures (Couper, 2013), which would facilitate the development of shorter questionnaires that do not include multi-item measures and are less burdensome for respondents.

Finally, there are some topics for which a rather vague research agenda can be set at the moment, primarily because we do not know what technological advances that can affect survey research will follow in the future. One example of this would be the unavailability of the complete list of email addresses of the general population from which samples can be drawn to conduct online (panel) surveys. Although it seems unlikely that a full list of email addresses will be formed any time soon or that it will be made available to researchers (if, for example, population registries are supplemented with email addresses), this does not necessarily mean that such a list will be unavailable in the future. This suggestion is speculative; however, as official institutions adopt technological innovations (e.g., digital transactions with citizens – e-government), certain groups might be reached for surveys in the future. For example, the German Postal Service has a service of digital mail,[36] which is not yet wide-spread. However, in analogy to the address-based sampling that has been increasingly adopted in the USA, a scenario is imaginable in which it would be possible to sample the users of such digital post services. Even if the full list of the population is not available, such a frame could be used as one of the multiple frames. Having multiple frames together with multiple modes has been long predicted by survey methodologists (Groves & Lyberg, 2010). In this case, it would be a challenge for survey statisticians to bring the frames together. To conclude, survey methodology is to a great extent affected by technological changes and by the societal acceptance of these changes. Similar to the manner in which research questions and practices concerning online panels benefit from the past research on mail surveys (since both types of surveys are self-administered), future mixed-mode, multiple-device, and multiple-frame surveys can benefit from the insight of the studies on probability-based online panel surveys.

[36] http://www.epost.de/privatkunden/e-mail-adresse-epostbrief.html

Appendix A (Appendix for Chapter 2)

Table A1. Logistic regression on very good/good health by reference surveys

	Odds Ratio	Standard error
Male	1.031	(0.072)
Age group		
18-24	Ref.	
25-34	0.681**	(0.094)
35-49	0.468***	(0.064)
50-64	0.267***	(0.037)
65+	0.358***	(0.058)
Education		
low	Ref.	
middle	1.331**	(0.124)
high	2.138***	(0.200)
Married	1.175*	(0.093)
Working	1.384***	(0.118)
Immigration background	1.097	(0.120)
ALLBUS	1.681***	(0.172)
ESS	1.418**	(0.143)
Constant	1.320	(0.213)
N	4600	
R^2	0.051	

$^{+}$ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.
All surveys combined, design weighted, GOPP is reference category.

Table A2. Variables for comparison from the Online Panel and mode they have been surveyed in.

Variable	Mode of data collection	Measured online in the first questionnaire
Gender	CATI recruitment	Yes
Year of birth/age	CATI recruitment	Yes
Educational attainment	CATI recruitment	No
Marital status	CATI recruitment	No
Employment status	CATI recruitment	No
Immigration background	CATI recruitment	No
Trust	CATI recruitment	Yes (different operationalization)
Political interest	Online first questionnaire	-
Health status	Online first questionnaire	-
Satisfaction with government	Online first questionnaire	-
Satisfaction with health services in Germany		
German economy	Online first questionnaire	-
German economy in a year	Online first questionnaire	-
Financial situation in household	Online first questionnaire	-
Financial situation in household in a year	Online first questionnaire	-

Calculation of the recruitment rate in GOPP:

For probability-based Internet panels, AAPOR (2011) provides the formulae to calculate the recruitment rate (RECR) and the profile rate (PROR), as well as the completion rate (COMR). These formulae were adopted from an earlier publication by Callegaro and DiSogra (2008).

The initial recruitment rate (RECR) can be computed as follows (AAPOR, 2011, p. 37):

$$RECR = \frac{IC}{IC+(R+NC+O)+e(UH+UO)}$$

Because the target population in GOPP is a subset of the full population, namely Internet users, it is necessary to multiply the number of non-interviews with the Internet use rate in our RDD sample to obtain the number of people who would be eligible for the panel. The Internet use rate is the proportion of respondents who use the Internet (n=3514) among all interviews (n=4840) and is equal to 0.726. Thus, the formula for RECR is:

$$RECR = \frac{IC}{IC+[(R+NC+O)+e(UH+UO)]IU} = \frac{1665}{1665+[(14467+3332+432)+0.285(13694+547)]0.726} = 9.3\%$$

For the estimation of *e*, we used a proportional estimation approach, which is based on the available data. The definition of e is the proportion of all known eligible cases to the sum of all known eligible and not eligible cases. We do not assume that unknown cases are 100% eligible, which would be similar to AAPOR RR1, yielding a maximum response rate. We further do not assume that these cases are 0% eligible, which would be similar to AAPOR RR 5, and resulting in a minimum response rate.

The formula for estimating *e* is:

$$e = \frac{eligible\ cases}{eligible\ cases + not\ eligible\ cases} = \frac{I+P+R+NC+O}{I+P+R+NC+O+NE}$$

Thus, e is calculated as follows:

$$e = \frac{eligible\ cases}{eligible\ cases + not\ eligible\ cases} = \frac{I+P+R+NC+O}{I+P+R+NC+O+NE} = \frac{4840+0+14467+3332+432}{4840+0+14467+3332+57903} = 0.285$$

The profile rate (PROR) and the completion rate (COMR) are calculated using the standard formulae in AAPOR (2011, p. 37).

$$PROR = \frac{(I+P)}{(I+P)+(R+NC+O)} = \frac{1010}{1010+655} = 60.7\%,$$

where I+P are all respondents who started the first online survey

&

$$COMR = \frac{(I+P)}{(I+P)+(R+NC+O)} = \frac{934}{1010+655} = 56.1\%,$$

where I+P are all respondents who completed the first online survey

Since in GOPP the first online survey is at the same time the profile survey, the cumulative response rate (CUMRR) for the first survey is:

$$CUMRR = RECR \times COMR = 9.3\% \times 56.1\% = 5.2\%$$

IC = Initial consent
I = Complete interview
P = Partial interview
R = Refusal and break-off
NC = Non-contact
O = Other
UH = Unknown if household/occupied
UO = Unknown other
e = Estimated proportion of cases of unknown eligibility that are eligible
NE = Not eligible
IU = Estimated proportion of Internet users

The final dispositions of case codes were coded using final disposition codes for RDD telephone surveys (AAPOR, 2011).

Appendix B (Appendix for Chapter 3)

Table B1. Descriptive statistics of variables used for multivariate analysis.

Variables	Mean & (SD) / Percent	N	Min	Max
Recruitment rate	51.36%	2656		
Started online survey (logged in)	60.53%	1310		
Response (excludes those who quit at the first page)		1283		
Age	43.26 (15.01)	2577	18	89
Sex (male)	51.06%	2634		
Education		2548		
Low (quit school without a diploma, lower vocational education)	16.05%			
Medium (higher vocational education)	33.87%			
High (high school diploma)	50.08%			
Still at school, other to missing				
Employment status (working=1: fill or part-time, nonworking=0: maternity leave, students, retired, unemployed & other)	73.46% 71.80%	1944 2532		
Immigration background (German=0 defined as born in Germany and has German citizenship)	14.81%	2634		
Married or has partner in hh	60.31%	2595		
Single household	22.06%	2602		
Children under 18 in household	33.60%	2598		
Goes out		2625		
every week	28.65%			
every month	30.86%			
less than once a month	34.06%			
never	6.44%			
Sports		2627		
every week	60.11%			
every month	7.16%			
less than once a month	16.56%			
never	16.18%			
Helps friends/family		2565		
every week	30.29%			
every month	40.31%			
less than once a month	26.51%			
never	2.88%			

Table B1. Continued.

Variables	Mean & (SD) / Percent	N	Min	Max
Trust		2619		
don't trust	26.04%			
trust	15.85%			
it depends	58.11%			
Incentive in Euros		2656		
Incentive=0	25.79%			
Incentive=2	26.32%			
Incentive=5	36.52%			
Incentive=10	11.37%			
Income question asked	49.68%	2639		
Refused to answer income question, including DK	20.06%	1326		
Urbanity (1=big city & suburb, 0=small town, village)	38.68%	1978		
Mobile interview (mobile)	41.64%	2656		
Has mobile phone (cell)	95.37%	2636		
Internet usage in years (those started in 2011 coded as 1 year)	9.48 (5.22)	2586	1	31
Frequency of Internet usage		2631		
daily	63.17%			
more than once a week	22.84%			
once a week	9.50%			
once a month	3.08%			
less than once a month	1.41%			
Participated online survey (recoded to binary 1=participated, 0=never participated)		2616		
once	10.13%			
more than once	10.44%			
never	79.43%			
Topic interest (topic named)	88.43%	1366		

Note: summary statistics for those who were asked the recruitment question

&

Table B2. An additional logistic regression model with the dependent variable "willing to participate in the online panel."

	Model 3b	
Variable	**Odds ratio**	**SE**
Demographics		
Age	0.985**	(0.004)
Gender (male)	1.182+	(0.108)
Education (low)	ref.	—
middle	1.216	(0.164)
high	1.307*	(0.175)
Working (yes)	0.773*	(0.079)
Immigrant (yes)	0.912	(0.118)
Married/partner	0.710**	(0.072)
Children <18	1.021	(0.106)
Social Integration		
Going out	0.948	(0.050)
Sport	0.997	(0.038)
Helping out	1.015	(0.056)
Trust (no)	ref.	—
yes	1.340*	(0.193)
it depends	1.295*	(0.139)
Survey Design		
Incentive in euros (none)	ref.	—
2 Euros	1.103	(0.134)
5 Euros	1.488**	(0.171)
10 Euros	2.718**	(0.486)
Fieldwork agency	0.740**	(0.072)
Survey specific characteristics		
Interviewed on cell	1.226*	(0.113)
Internet usage, years	1.028**	(0.009)
Frequency of Internet use	excluded	—
Online survey experience	2.283**	(0.263)
Constant	1.171	(0.423)
Sample size	2300	
Pseudo R^2	0.07	
BIC	3112.222	

Note: + p<0.10, * p<0.05, ** p<0.01, unweighted data.

Appendix C (Appendix for Chapter 4)

Table C1. Descriptive statistics by wave.

	Wave 1	N	Wave 2	N	Wave 3	N	Wave 4	N	Wave 5	N	Wave 6	N	Wave 7	N	Wave 8	N
Wave-response	56.30	1643	48.93	1643	47.60	1643	46.68	1643	45.16	1643	42.60	1643	43.52	1643	40.72	1643
Started (if invited)	60.86	1643	51.89	1636	50.19	1618	49.34	1593	47.53	1578	46.34	1569	46.54	1560	45.76	1556
Completed (if started)	92.50	1000	94.70	849	96.31	812	97.58	786	98.93	750	96.29	727	98.48	726	95.79	712
Item- nonresponse, %	.398	925	.676	804	.438	782	.159	767	.153	742	.199	700	.326	715	.225	682
Duration, min.	13.394 (5.591)	778	13.542 (6.618)	683	11.420 (5.623)	676	10.384 (4.938)	686	12.424 (5.772)	632	13.090 (5.566)	603	11.378 (5.539)	615	21.632 (10.410)	485
Age	42.689 (14.561)	924	43.177 (14.622)	791	43.065 (14.562)	769	43.423 (14.677)	759	43.465 (14.710)	733	43.415 (14.914)	690	43.562 (14.719)	705	43.521 (14.741)	660
Male	53.51	925	53.11	804	51.41	782	52.41	767	50.94	742	51.86	363	52.17	715	52.47	669
Education:																
low	11.12		11.04	779	9.95	764	10.28	749	9.25	724	9.52	683	9.20	696	9.45	667
middle	28.33		27.86		28.53		26.70		28.31		27.67		27.30		27.89	
high	60.55		61.10		61.52		63.02		62.43		62.81		63.51		62.67	
In paid work	71.82	880	73.19	787	72.08	770	71.31	753	71.47	729	71.03	687	71.71	700	71.83	671
Immigration background	9.77	768	10.52	675	9.42	658	9.52	651	9.37	630	9.20	598	9.56	607	9.62	582
Married/has partner	69.83	812	68.98	719	69.01	710	68.95	702	68.54	677	68.66	635	68.98	648	69.08	621
Incentive: 0	17.30	925	16.79	804	15.86	782	16.43	767	15.63	742	15.71	700	15.10	715	15.25	682
Incentive: 2	24.22		23.38		23.53		22.69		23.72		22.71		23.22		22.73	
Incentive: 5	24.65		24.00		24.55		24.64		23.85		24.29		24.06		24.05	
Incentive: 5+20	17.41		18.03		18.41		18.38		19.00		18.71		19.02		19.21	
Incentive: 10+20	16.43		17.79		17.65		17.86		17.79		18.57		18.60		18.77	
Overall evaluation	3.746 (.597)	924	3.616 (.698)	794	3.732 (.663)	776	3.764 (.717)	767	3.747 (.711)	742	3.402 (.897)	699	3.628 (.807)	715	3.758 (.724)	682

Survey evaluation:																
interesting	2.978 (.625)	911	2.948 (.706)	800	3.015 (.685)	777	3.037 (.714)	766	3.067 (.704)	740	2.805 (.799)	699	3.014 (.754)	714	3.196 (.683)	680
diverse	2.996 (.565)	915	2.745 (.724)	796	2.798 (.687)	779	2.890 (.681)	765	2.923 (.671)	739	2.555 (.840)	699	2.790 (.709)	714	3.193 (.662)	680
important for science	2.815 (.660)	895	2.782 (.693)	790	2.876 (.665)	775	2.876 (.727)	764	2.912 (.714)	738	2.547 (.792)	696	2.847 (.754)	712	2.820 (.707)	678
long	1.829 (.660)	907	1.864 (.712)	797	1.742 (.635)	778	1.673 (.593)	765	1.745 (.671)	738	2.000 (.814)	699	1.724 (.602)	715	2.250 (.850)	681
difficult	1.408 (.545)	904	1.503 (.599)	797	1.486 (.599)	778	1.563 (.657)	766	1.759 (.741)	739	1.597 (.662)	699	1.948 (.812)	715	2.073 (.777)	681
too personal	1.695 (.634)	909	2.205 (.811)	801	2.160 (.823)	777	1.919 (.768)	765	1.804 (.718)	739	2.451 (.940)	700	2.062 (.896)	715	2.222 (.861)	681

Note: means and percentages are reported, standard deviation in parentheses. Due to the changed order of administering questionnaires, survey evaluation items reflect content evaluation of questionnaire content only for waves 1, 7, and 8. INR is short for item nonresponse.

Table C2. Logistic regression: dependent variable "ever redeemed incentive".

	OR	SE	p-level
M: interesting	.891	.472	0.827
M: diverse	.404	.194	0.059
M: important for science	1.924	.597	0.035
M: long	1.369	.528	0.415
M: difficult	.929	.309	0.824
M: too personal	1.347	.385	0.299
M: overall	2.687	1.257	0.035
M: INR	.001	.001	0.597
Age	.958	.009	0.000
Male	.803	.197	0.371
Education: low			
Education: middle	1.609	.636	0.228
Education: high	2.234	.830	0.031
In paid work	.922	.247	0.762
Immigration background	.520	.213	0.111
Married/has partner	1.030	.272	0.910
Incentive: 2	reference		
Incentive:5	2.488	.782	0.004
Incentive: 5+20	2.299	.735	0.009
Incentive: 10+20	4.083	1.426	0.000
Completed surveys	2.983	.314	0.000
Constant	.000	.000	0.000
N	669		
Log likelihood	-237.536		
Pseudo R^2	0.476		

Table C3. Hybrid model: interaction incentive amount and incentive redemption.

	OR	SE	p-level
D: interesting	1.135	.218	0.510
D: diverse	.953	.178	0.796
D: important for science	.753	.124	0.086
D: long	.875	.151	0.441
D: difficult	1.462	.254	0.029
D: too personal	1.621	.241	0.001
D: overall	1.089	.199	0.642
M: interesting	1.122	.497	0.795

Table C3. Continued

	OR	SE	p-level
M: diverse	.637	.264	0.276
M: important for science	.785	.232	0.410
M: long	2.373	.784	0.009
M: difficult	.479	.155	0.023
M: too personal	.865	.227	0.579
M: overall	1.139	.447	0.740
Age	.964	.009	0.000
Male	1.000	.256	1.000
Education: low	reference		
Education: middle	.729	.292	0.431
Education: high	.638	.240	0.234
In paid work	1.058	.297	0.840
Immigration background	1.032	.420	0.939
Married/has partner	1.102	.306	0.727
Incentive: 2	reference		
Incentive:5	1.376	.517	0.396
Incentive: 5+20	.698	.292	0.390
Incentive: 10+20	1.082	.495	0.862
Payment	.0144	.008	0.000
Payment*incentive 5	.642	.449	0.526
Payment*incentive 5+20	.780	.619	0.754
Payment*incentive 10+20	.563	.434	0.456
Wave 2	reference		
Wave 3	.523	.156	0.029
Wave 4	.616	.183	0.103
Wave 5	.797	.237	0.445
Wave 6	1.289	.384	0.394
Wave 7	.545	.200	0.099
Wave 8	.897	.294	0.739
Constant	7.176	12.336	0.252
N observations	3764		
N persons	669		
lnsig2u	1.180	.239	
sigma_u	1.804	.215	
Rho	.497	.060	
Wald χ^2	176.89		0.000

Note: Log likelihood -714.198.

Table C4. Hybrid model: interaction survey experience and incentive redemption.

	OR	SE	p-level
D: interesting	1.011	.223	0.961
D: diverse	.972	.211	0.897
D: important for science	.653	.125	0.027
D: long	1.002	.202	0.991
D: difficult	1.573	.323	0.028
D: too personal	1.791	.317	0.001
D: overall	1.023	.226	0.919
M: interesting	1.118	.504	0.805
M: diverse	.638	.268	0.285
M: important for science	.783	.235	0.414
M: long	2.451	.819	0.007
M: difficult	.467	.153	0.020
M: too personal	.863	.230	0.580
M: overall	1.171	.467	0.693
Age	.964	.010	0.000
Male	.995	.260	0.984
Education: low	reference		
Education: middle	.742	.302	0.463
Education: high	.641	.246	0.247
In paid work	1.061	.302	0.835
Immigration background	1.009	.417	0.983
Married/has partner	1.122	.317	0.684
Incentive: 2	reference		
Incentive: 5	1.227	.395	0.524
Incentive: 5+20	.659	.241	0.253
Incentive: 10+20	.882	.326	0.733
Payment	0.009	.004	0.000
Payment*D: interesting	1.815	.876	0.217
Payment*D: diverse	.937	.407	0.882
Payment*D: important for science	1.706	.665	0.171
Payment*D: long	.563	.226	0.153
Payment*D: difficult	.776	.297	0.507
Payment*D: too personal	.777	.243	0.420
Payment*D: overall	1.104	.457	0.810
Wave 2	reference		
Wave 3	.498	.151	0.021
Wave 4	.607	.182	0.096

Table C4. Continued

	OR	SE	p-level
Wave 5	.777	.233	0.401
Wave 6	1.260	.378	0.441
Wave 7	.523	.196	0.084
Wave 8	.881	.291	0.702
Constant	7.021	12.241	0.264
N observations	3764		
N persons	669		
lnsig2u	1.237	.221	
sigma_u	1.857	.256	
Rho	.512	.059	
Wald χ^2	175.35		0.000

Note: Log likelihood: -708.325.

&

Table C5. Logistic regression: dependent variable "stay after wave 8".

	OR	SE	p-level
M: interesting	.786	.521	0.716
M: diverse	1.116	.638	0.847
M: important for science	1.340	.523	0.454
M: long	.477	.226	0.119
M: difficult	.977	.418	0.957
M: too personal	.660	.230	0.233
M: overall	4.587	2.583	0.007
Age	.967	.011	0.002
Male	1.606	.501	0.128
Education: low	reference		
Education: middle	.998	.512	0.997
Education: high	1.534	.746	0.379
In paid work	.820	.278	0.559
Immigration background	.682	.350	0.456
Married/has partner	1.166	.387	0.644
Incentive: yes	2.968	1.071	0.003
Completed surveys	1.104	.148	0.458
Constant	.147	.344	0.411
N	560		
Log likelihood	-162.600		
Pseudo R^2	0.183		

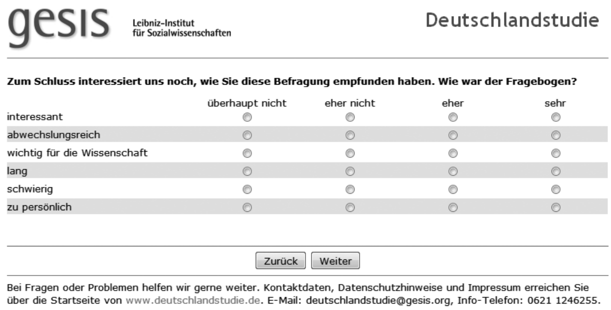


Figure C1. The survey evaluation question.
Question text: Finally, we are interested how do you feel about the questionnaire. How was the questionnaire? Items: “interesting”, “diverse” (the German word “abwechslungreich” can be translated as rich in variety, multifaceted, diverse; throughout the dissertation “diverse” is used), “important for science”, “long”, “difficult”, “too personal”. The scale categories: “not at all”, “rather not”, “rather”, “very”.

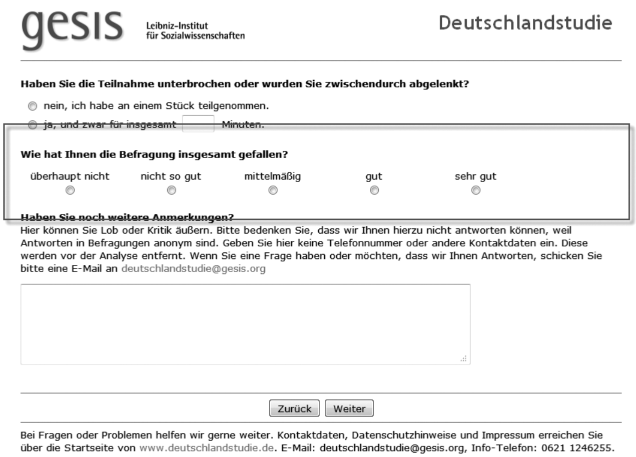
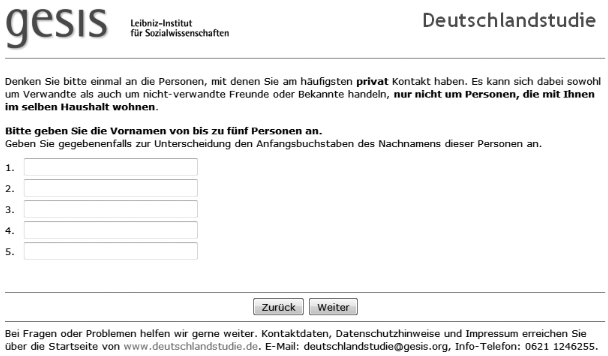


Figure C2. The overall evaluation of the questionnaire.

Question text: Overall, how did you like the survey? Scale: “not at all”, “not so good”, “moderately”, “good”, “very good”.

Appendix D (Appendix for Chapter 5)



gesis Leibniz-Institut für Sozialwissenschaften

Deutschlandstudie

Denken Sie bitte einmal an die Personen, mit denen Sie am häufigsten **privat** Kontakt haben. Es kann sich dabei sowohl um Verwandte als auch um nicht-verwandte Freunde oder Bekannte handeln, **nur nicht um Personen, die mit Ihnen im selben Haushalt wohnen**.

Bitte geben Sie die Vornamen von bis zu fünf Personen an.
Geben Sie gegebenenfalls zur Unterscheidung den Anfangsbuchstaben des Nachnamens dieser Personen an.

1.
2.
3.
4.
5.

Zurück Weiter

Bei Fragen oder Problemen helfen wir gerne weiter. Kontaktdaten, Datenschutzhinweise und Impressum erreichen Sie über die Startseite von www.deutschlandstudie.de. E-Mail: deutschlandstudie@gesis.org, Info-Telefon: 0621 1246255.

Figure D1. Screenshot of the name generator question

Text of the names generator question: please think about persons with whom you often **privately** have contact. They could be relatives and persons not related to you, friends or acquaintances, but not persons who **live with you in the same household.**
Please provide the first names of up to five persons. If needed, please type in the first letter of the surname of these persons to differente between them.

Note for the Table 5.4

True distributions for the knowledge question for the time questions were fielded in 2011:

Item 1: The EU has the largest number of commercial nuclear power stations (for electricity production) in the world. True, combined as of July 2011 the EU had the largest number of power plants in the world followed by the USA (World Nuclear Association, 2011b).

Item 2: Nuclear power plants are the only producers of radioactive waste. False, low-level radioactive waste is also produced by hospitals, laboratories and industry (World Nuclear Association, 2012).

Item 3: About a third of the electricity produced within the EU is produced by nuclear power plants. True, as of April 2011, about 31 percent of the EU's energy came from nuclear power plants (Stull, 2011).

Item 4: New nuclear power plants are presently being constructed in Germany at this very moment. False, the questionnaire was fielded in August 2011 following the Fukushima nuclear disaster, at that time and after that nuclear power plants in Germany were shutting down (World Nuclear Association, 2011a).

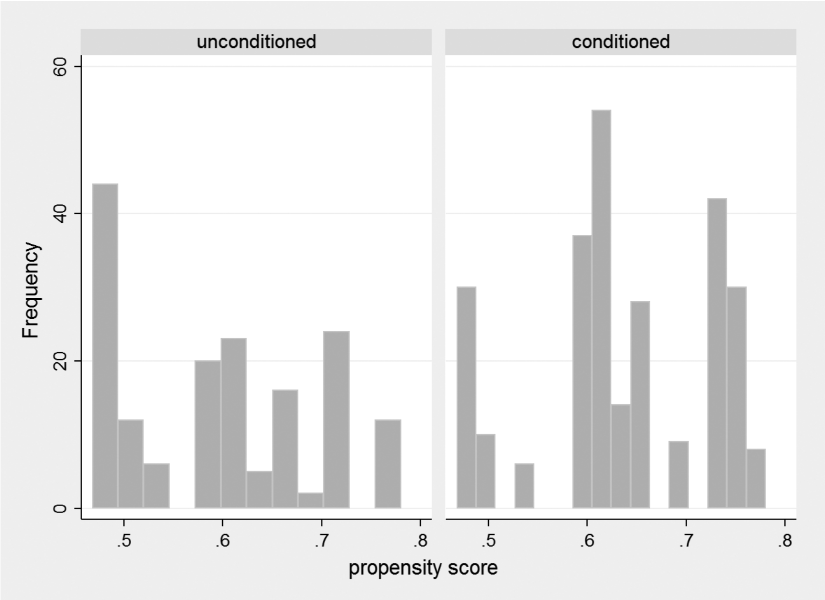


Figure D2. Distributions of propensity scores between the two experimental groups.

Results of the factor analysis for Hypothesis 2 with the ISSP 2010 and the GOPP data: in both surveys and for ISSP 2010 for both groups of Internet users and non-users all items load on one factor, primary factor loadings are all over 0.4. For all analyses principal components factor analysis was performed.

&

Table D1. Factor loadings, uniqueness, explained variance, sample sizes and eigenvalues

Items environmental behavior	ISSP 2010 All		ISSP 2010 Internet users		GOPP	
	Factor loading	**Uniqueness**	**Factor loading**	**Uniqueness**	**Factor loading**	**Uniqueness**
special effort to sort glass or tins or plastic or newspapers and so on for recycling	0.4518	0.7958	0.4763	0.7731	0.5463	0.7015
special effort to buy fruit and vegetables grown without pesticides or chemicals	0.5647	0.6811	0.5823	0.6609	0.6824	0.5344
cut back on driving a car for environmental reasons	0.6007	0.6392	0.6069	0.6317	0.6184	0.6176
reduce the energy or fuel use at home for environmental reasons	0.7627	0.4183	0.7662	0.4130	0.7540	0.4315
save or re-use water for environmental reasons	0.7586	0.4246	0.7525	0.4337	0.7314	0.4650
avoid buying certain products for environmental reasons	0.7693	0.4081	0.7696	0.4077	0.7581	0.4253
N	1146	828	412			
Eigenvalue	2.633	2.680	2.825			
Percent of variance explained	43.88	44.67	47.08			

Table D2. Descriptive statistics for demographic and attitudinal variables for the two experimental groups from the second quasi-experiment with means (M) or proportions (%), standard deviations (SD), sample sizes (N), and tests for statistically significant differences.

	Unconditioned (T=0)			Conditioned (T=1)			Difference	
Variable	**M/%**	**SD**	**N**	**M/%**	**SD**	**N**	**Test statistic**	***p*-level**
Male	55.92%	.498	211	53.93%	.499	267	$\chi^2(1)= 0.188$	.664
Age	44.41	15.013	206	43.01	14.330	266	t=1.028	.304
Education		.717	199		.670	263	$\chi^2(2)= 1.210$	.546
low	13.07%			9.89%				
middle	26.13%			28.14%				
high	60.80%			61.98%				
In paid work	75.00%	.434	200	70.45%	.457	264	$\chi^2(1)=1.177$	.278
Trust	2.66	.595	185	2.78	.652	256	t=-1.854	.064
Political interest	2.72	.694	187	2.67	.782	258	t=0.992	.474
Self-assessed health	3.58	.891	187	3.68	.925	258	t=0.464	.232
Survey participation in the past 12 months	46.24%	.499	186	38.37%	.487	258	$\chi^2(1)=2.750$	.097
Overall evaluation Q1	3.71	.643	187	3.79	.589	258	t=-1.376	.169
Q1 interesting	78.69%	.787	183	84.50%	.845	258	$\chi^2(1)=2.456$	.117
Q1 varied	83.33%	.374	186	84.44%	.363	257	$\chi^2(1)=0.097$	.755
Q1 important f. science	67.21%	.471	183	75.49%	.431	253	$\chi^2(1)=3.614$	.057
Q1 long	9.34%	.292	182	10.55%	.308	256	$\chi^2(1)=0.171$	.679
Q1 difficult	1.64%	.127	183	0.78%	.088	257	$\chi^2(1)=0.706$	.401
Q1 too personal	8.15%	.274	184	5.06%	.232	257	$\chi^2(1)=1.726$	.189
Incentive		.818	211		.501	267	--	--
0	29.38%						--	--
2	32.70%						--	--
5/5+bonus	37.91%			50.94%			--	--
10+bonus				49.06%				

Note: sample sizes vary due to missing values on variables.

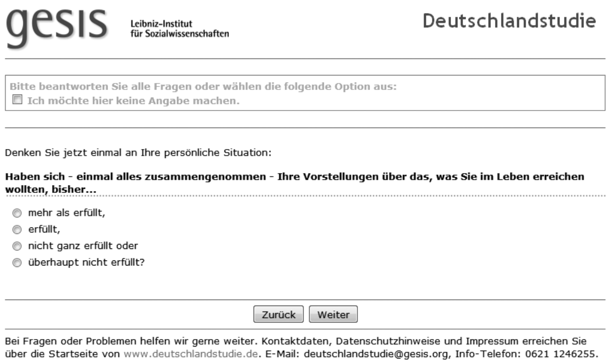


Figure D3. Implementation of the edit-checks in the GESIS Online Panel Pilot

Translation of the edit-check: Please answer all the questions or choose the following option: [checkbox] I do not want to provide an answer here.

Appendix E (Appendix for Chapter 6)

Table E1. Distributions of the demographic variables used for weighting.

	GOPP Questionnaires 1-5 & 7			ALLBUS 2010 Internet users			ALLBUS 2012 Internet users		
Variable	**%**	**SE**	**CI**	**%**	**SE**	**CI**	**%**	**SE**	**CI**
Male	53.1	.015	[50.1, 56.0]	51.5	.012	[49.3, 53.8]	51.6	.010	[46.4, 50.3]
Age groups									
18–24	13.0	.010	[11.2, 15.1]	13.9	.008	[12.4, 15.6]	13.9	.007	[12.6, 15.3]
25–34	19.8	.012	[17.6, 22.3]	18.6	.009	[16.9, 20.4]	17.4	.008	[15.9, 18.9]
35–49	34.9	.014	[32.2, 37.8]	35.4	.011	[33.2, 37.6]	31.4[a]	.009	[29.6, 33.3]
50–64	23.4	.013	[21.0, 26.0]	23.2	.010	[21.4, 25.2]	27.7[a]	.009	[26.0, 29.5]
65+	8.8	.008	[7.3, 10.6]	8.9	.007	[7.6, 10.2]	9.6	.006	[8.5, 10.9]
Education									
low	12.5	.010	[10.7, 14.6]	21.4	.010	[19.6, 23.4]	20.9	.008	[19.4, 22.6]
medium	30.0	.014	[27.4, 32.7]	37.8	.011	[35.6, 40.0]	39.4	.010	[37.5, 41.3]
high	57.5	.015	[54.6, 60.4]	40.8	.011	[38.5, 43.0]	39.7	.010	[37.8, 41.6]

[a] significant differences ($p<0.01$) between ALLBUS 2010 and 2012. N for GOPP is 1114 respondents, which includes the respondents who completed at least one of the questionnaires 1 to 5, or 7. SE is short for standard error, CI is short for confidence interval.

&

Table E2. Distributions of demographic characteristics in the GOPP over questionnaires (percentages).

Variable/Q	1	2	3	4	5	7
Male	52.97 (.016) [49.9, 56.1]	52.74 (.017) [49.4, 56.1]	51.75 (.018) [48.3, 55.2]	51.48 (.018) [48.0, 55.0]	51.91 (.018) [48.4, 55.5]	51.99 (.019) [48.4, 55.6]
Age groups						
18–24	13.37 (.011) [11.4, 15.6]	12.53 (.011) [10.5, 15.0]	12.12 (.012) [10.0, 14.6]	12.52 (.012) [10.4, 15.0]	13.27 (.012) [11.0, 15.9]	12.35 (.012) [10.1, 14.9]
25–34	19.31 (.012) [17.0, 21.9]	19.21 (.014) [16.7, 22.0]	19.25 (.014) [16.7, 22.1]	18.45 (.014) [15.9, 21.3]	18.53 (.014) [15.9, 21.5]	19.20 (.015) [16.5, 22.2]
35–49	34.85 (.015) [32.0, 37.8]	34.73 (.016) [31.6, 38.0]	35.00 (.017) [31.8, 38.3]	35.10 (.017) [31.8, 38.5]	34.03 (.017) [30.7, 37.5]	34.29 (.018) [30.9, 37.8]
50–64	23.47 (.013) [20.9, 26.2]	24.22 (.015) [21.4, 27.2]	24.50 (.015) [21.6, 27.6]	24.65 (.015) [21.7, 27.8]	24.97 (.017) [22.0, 28.2]	24.97 (.016) [22.0, 28.2]
65+	9.00 (.009) [7.4, 10.9]	9.31 (.010) [7.5, 11.5]	9.13 (.010) [7.3, 11.3]	9.28 (.010) [7.4, 11.6]	9.20 (.010) [7.3, 11.5]	9.19 (.011) [7.3, 11.5]
Education						
low	12.18 (.010) [10.3, 14.3]	11.81 (.011) [9.8, 14.2]	10.25 (.011) [8.3, 12.6]	10.06 (.011) [8.1, 12.4]	10.25 (.011) [8.3, 12.6]	10.01 (.011) [8.0, 12.4]
medium	29.50 (.014) [26.8, 32.4]	29.00 (.016) [26.0, 32.2]	28.00 (.016) [25.0, 31.2]	28.52 (.016) [25.4, 31.8]	27.73 (.016) [24.7, 31.0]	27.57 (.017) [24.4, 30.9]
high	58.32 (.016) [55.2, 61.3]	59.19 (.017) [55.8, 62.5]	61.75 (.017) [58.3, 65.1]	61.42 (.017) [57.9, 64.8]	62.02 (.018) [58.5, 65.4]	62.41 (.018) [58.8, 65.9]
N	1010	838	800	775	761	729

Note: standard errors in parentheses, 95%-confidence intervals in brackets. The data are not weighted. Q is short for GOPP questionnaire, SE is short for standard error, CI is short for confidence interval.

Table E3. Post-stratification weights obtained with IPF-Weighting.

Questionnaire	1	2	3	4	5	7
N	1010	838	800	775	761	729
Min	0.585	0.539	0.520	0.510	0.507	0.498
Max	2.299	2.367	2.445	3.268	2.956	3.086
1% percentile	0.585	0.539	0.520	0.510	0.507	0.498
99% percentile	2.044	2.137	2.907	2.514	2.528	2.558

&

Calculation of the effect sizes

Formulas used for calculating the weighted effect sizes:

$$\overline{ES} = \frac{\sum_{i=1}^{k}(w_i \times ES_i)}{\sum_{i=1}^{k} w_i};\ w_i = \frac{1}{SE_i^2}$$

where *w* is the inverse variance weight, *ES* is the unweighted effect size *(d)*, and *SE* is the standard error of the difference between the effect sizes. For the calculation of the effect sizes the effect size calculator at http://www.campbellcollaboration.org/resources/effect_size_input.php was used. The formulas for calculating the effect sizes can be found in (Lipsey & Wilson, 2001, p. 172 ff.). We used the unweighted sample sizes to calculate the mean effects.

Questions used in the analysis

Asterisk marks the questions where a card with answer options was offered to the respondents by interviewers in ALLBUS interviews. Question wordings are provided as translated by ALLBUS team, original German question wordings were identical among surveys unless otherwise noted.

German economy*
How would you generally rate the current economic situation in Germany?

Very good
Good
Partly good/partly bad
Bad
Very bad

Own financial situation*
And your own current financial situation?

Very good
Good
Partly good/partly bad
Bad
Very bad

German economy in 1 year*
What do you think the economic situation in Germany will be like in one year?

Considerably better than today
Somewhat better than today
The same
Somewhat worse than today
Considerably worse than today

Own financial situation in 1 year*
And what will your own financial situation be like in one year?

Considerably better than today
Somewhat better than today
The same
Somewhat worse than today
Considerably worse than today

General health*
A question about your health: How would you describe your health in general?

Very good
Good
Satisfactory
Poor
Bad

Employment status*
And now let's continue with employment and your occupation. Which of the categories on the card applies to you?

Full time employment
Part ("half") time employment
Less than part ("half") time employment
Not working

If there are difficulties referring to the classification, here are some hints for you:
Trainees are considered employees in a regular occupation.
Family members assisting in a family business who are full- time or part- time ("halftime") employees in the business of a household or a family member, without having a formal contract, are also considered employees in a regular occupation (either full- time or part- time).

"Employed less than part- time" are persons who are gainfully employed while, at the same time, one of the following applies:
Attend a full- time school (pupils and students),
Are registered as unemployed or
Draw a retirement benefit / pension as a result of previous employment.
Persons on maternity / parental leave or on another type of leave of absence are not considered employees in a regular occupation.

Marital status*
What is your marital status? Are you...

Married and living with your spouse
Married and living apart
Widowed
Divorced
Never married

Civil partnership, living together
Civil partnership, living apart
Registered partner deceased
Civil partnership dissolved

Note that in GOPP this question involved a filter: "civil partnership" category was added to the first five answer options and followed by the four civil partnership answer choices in case a respondent selected the category "civil partnership".

&

Religiosity*
Would you describe yourself as tending to be religious or tending to be not religious?
(1) "religious"
(10) "not religious"

Note: 10pt-scale, in ALLBUS end labels "religious", "not religious" and random letters for scale points, in GOPP end labels only.

Confession*
May I ask what religious confession you belong to?

The Roman Catholic church
The German Protestant church (excluding free churches)
A Protestant free church
Another Christian denomination
Another non-Christian religion
No religious affiliation

Note: In GOPP "May we ask what religious confession you belong to?"

Frequency of church attendance
As a rule, how often do you go to church?

More than once a week
Once a week
Between one and three times a month
Several times a year
Less
Never

Note: in ALLBUS 2012 the question is split into two: "How often go to church?" for respondents who belong to a Christian religious denomination and "How often go you go to church or to mosque, synagogue or other chapel?"

Born in Germany
Were you born within the current borders of Germany?

Yes
No

Citizenship
What citizenship do you have? If you have several citizenships, please name all of them.

Germany
Greece
Italy
Former Yugoslavia:
Bosnia and Herzegovina,
Serbia,
Croatia,
Macedonia,
Slovenia
Poland
Turkey
Other country, please enter ________________________
None, stateless

Self-assessed social class
There is a lot of talk about social class these days. What class would you describe yourself as belonging to?

Lower class
Working class
Middle class
Upper middle class
Upper class

Ownership of dwelling
The next question deals with the accommodation you/your family live in. Please tell me which of the categories on the card applies to you/your family.

Sublet
In an official/company flat
In subsidized municipal housing
In a rented flat (not subsidized housing)
In a rented house (detached/semi-detached)

In a flat owned by you or your family
In a house owned by you or your family
Other type of accommodation, please enter ________________________

Attitudes
I'm going to read you some statements now. Please tell me after each one whether you have the same or a different opinion.

No matter what some people say, life for ordinary people is getting worse rather than better.
With the future looking as it does, it's almost irresponsible to bring children into the world.
Most politicians are not really interested at all in the problems of ordinary people.
Most people don't really care in the slightest what happens to others.

Have the same opinion
Have a different opinion

Note: question wording for GOPP: "Please indicate for each statement whether you have the same or a different opinion."

&

Left-right orientation*
Many people use the terms "left" and "right" when they want to describe different political views. Here we have a scale which runs from left to right. Thinking of your own political views, where would you place these on this scale?

"left"
(10) "right"

Note: 10pt-scale, in ALLBUS end labels "left", "right" and random letters for scale points, in GOPP end labels only. The text "Here we have a scale which runs from left to right." omitted in GOPP.

Appendix F

A review of studies on panel conditioning according to their designs

In Chapter 5, it has been stated that somewhat mixed empirical evidence of occurrence, direction, and magnitude of panel conditioning in previous studies can be attributed to the absence of the theoretical framework and to the designs which are used to study the phenomenon. The goal of this excursus is to examine the latter part of this explanation by providing the overview of the studies on respondent conditioning according to their design. The classification further includes a distinction of a concept described in the prevous chapter of advantageous and disadvantageous conditioning.

Literature search

The initial literature search was performed via *ISI Web of Knowledge* (http://isiwebofknowledge.com) and *Google Scholar* (http://scholar.google.com) to account for grey literature. The search using keywords *panel* and *conditioning* resulted in 2393 journal articles on Web of Knowledge and 642000 entries on Google Scholar. The restriction to the phrase *panel conditioning* yielded 11 and 751 results proving this choice of keywords to be too restrictive. The initial search was refined using the words *panel, conditioning*, and *survey*. After searching the literature for keywords, the following phrases were identified to extend the scope of the search: *panel bias, time in sample, rotation group bias, re-interview effect, respondent conditioning, mere measurement effect,* and *measurement reactivity*. In addition, relevant bibliography references of reviews on panel conditioning were included into the set. Information on few further studies was obtained through personal contact to the researchers in the field. The resulting set contained 177 sources which seemed relevant after reading the abstracts and included journal articles, book chapters, reports, and conference presentations. The following sources were excluded from this overview:

- reviews and texts without empirical evidence or comments on earlier articles
- articles reporting the same study or duplicate publications
- studies without a description of the sample
- studies exploring learning in the course of a single interview
- studies with a self-prediction manipulation is, that is, both groups are interviewed, but one group receives an additional question (e.g. predict own cheating on the test)
- studies which concentrate solely on recall effect
- clinical trials with the dependent variables being occurrence/reporting of symptoms or specific medical studies (e.g., sending testing kits to respondents)
- studies using qualitative methods (e.g., respondents reviewing their survey experience)
- studies which focus on induced deliberation
- papers without a focus on panel conditioning.

Of all initially selected sources, 10 could not be obtained. Applying the eligibility criteria to the obtained sources led to a set of 89 sources.

Dimensions for the classification of studies

As was argued in Chapter 5, the discussion which sees panel conditioning exclusively as a negative phenomenon is not comprehensive. Thus, the study classification distinguishes between advantageous and disadvantageous panel conditioning. Further, it seems reasonable to separate correlational studies from manipulation studies in order to gain insight into the thesis mentioned at the beginning of this Excursus.

The last dimension takes into account the data available to the researchers: whether the true value is known or whether the direction of change due to panel conditioning can be specified. Some previous studies of panel conditioning has distinguished between question types –behavior, attitudes, and knowledge (e.g., Toepoel, Das, & van Soest, 2009; Yan, Datta, & Hepburn, 2011). This third dimension of knowing the true value or being able to predict the direction of change allows to take into account such differentiation by the question type. For knowledge and behavior question types, the external data may provide true values, which allow uncovering panel conditioning. Attitudinal questions generally allow only hypotheses with no direction. However, there are special cases like attitudes prone to social desirability so that a direction may be predicted. These considerations lead to the following classification:

1. Advantageous vs. disadvantageous panel conditioning
2. Correlational vs. manipulation methods of investigation
3. Knowledge of the true value: true value known, direction specified, or direction unspecified.

The dimension of panel of panel conditioning being disadvantageous or advantageous is associated either with a decrease in the data quality by producing measurement error (disadvantageous) or with improving the data quality if the reported answer is closer to the "true" value (advantageous).

Panel conditioning may refer either to changes in reporting the answers (as was the subject of studies in Chapter 5) or to changes of attitudes and behavior. In his review, Cantor (2008) classifies 20 selected studies on panel conditioning into three groups according to the outcomes: 1) changes in behavior, 2) changes in reporting behavior and 3) conditioning of attitudes and opinions where no distinction between changes in reporting and actual changes is possible. All changes in behavior, attitudes, and knowledge are classified as disadvantageous. Though increased vaccination (Battaglia et al., 1996) or increased blood donation (van Dongen, Abraham, Ruiter, & Veldhuizen, 2013) as a result of being interviewed can be seen as advantageous from an individual health perspective, these changes are disadvantageous in the sense of generalizability of findings to the population under study. Changes in reporting may be advantageous (improved reporting) and disadvantageous (misreporting).

The second dimension reduces the methods to two categories: studies which include manipulation (true experiments, field experiments, rotating panel design) and correlational studies (comparison with an unrelated cross-section or unrelated panel and various comparisons, which do not involve rotating design). The last dimension is the knowledge about the true value. Under the first category "true value is known" knowledge questions and questions on behavior would fall. The second category includes behavioral questions, knowledge questions asked in the form “Do you know…?” requiring the answer yes or no, and attitudes when a direction is given by social desirability. The third category includes attitudinal questions.

Results

The selected 89 studies were classified according to the dimensions identified above. Table F1 presents the aggregated results of the review (full references can be found in Table F2 at the end of this overview). In each cell the number of studies is provided. Additionally, the first number in parentheses indicates the number of studies with a defined population (probability-based). The second number is the number of population studies, which have 3 or more waves or are carried out in panels with 3 or more waves. Several conclusions can be inferred from Table F1.

First, disadvantageous panel conditioning received much more scholarly attention than advantageous panel conditioning. It should be noted that in this small number of studies not all of them were set to investigate the advantageous panel conditioning. Some studies found advantageous conditioning post-hoc when expecting to find disadvantageous conditioning. Second, there are about twice as many correlational studies as manipulation ones. Since studies which employed rotating panel design[37] were also classified as manipulation studies, thus experimental studies are few. Third, studies with a predicted direction or where no hypotheses about the direction are made outnumber those studies where outcomes are compared with a true value.

The design options represented by the cells of Table F1 are not equal. Some allow solid conclusions whereas others do not. Studies which allow comparisons with a true value allow separating true changes from changing in reporting. However, from all the study designs investigating disadvantageous conditioning (studies on advantageous conditioning are too few) this option has the least number of studies. The reason for that is that validation data are costly, difficult to obtain and absent for a substantial share of survey outcomes. Thus, manipulation with a clear direction is the viable option for most researchers. Correlational studies with a clear direction or studies where no direction is possible by far exceed other studies. This type of studies faces the problem of confounding the true changes in attitudes or behavior in the target population (in the society) with panel conditioning. Further conclusion of this overview is that the number of studies on panel conditioning based on a probability sample decreases with the favorability of the design. This is exacerbated for studies with probability-based samples and carried out in a setting of a panel (3 waves or more, see Table F2).

Conclusion

Overall, advantageous panel conditioning has received less scholarly attention than disadvantageous conditioning. The correlational studies prevail over manipulation studies. Studies which can make use of true values are rare, especially those carried out in panels which are based on a probability sample and have three or more measurement points.

[37] See Section 5.2 in Chapter 5 for a description of the design options.

Table F1. Classification of studies by effects on data quality, study designs, and availability of the true value.

	Disadvantageous		Advantageous	
	Manipulation	**Correlational**	**Manipulation**	**Correlational**
True value known	9 (5/1)	10 (9/6)	0	1 (1/1)
Direction clear	12 (9/6)	35 (26/22)	2 (2/2)	0
Direction unclear	11 (8/8)	24 (20/18)	0	3 (3/3)

Note: double counts are possible if studies fall into separate categories.

The two limitations of this overview are that it did not include the findings of studies or the effect sizes. This should be addressed in a formal meta-analysis of panel conditioning studies. Furthermore, design decisions such as conducting a correlational study or a manipulation study may be hard to influence in a practical setting and validation data may not be available. However, the option of analyzing more panel waves may be available to the researchers and should be considered in future studies. Additionally, the field could benefit if future studies differentiate between advantageous and disadvantageous panel conditioning and their opposing effects on data quality.

Table F2. Overview of the studies classified according to the dimensions

	Disadvantageous, Manipulation	Disadvantageous, Correlational	Advantageous, Manipulation	Advantageous, Correlational
True value known	Dholakia2010 Godin2008 *Kraut1973* Lefevere2010 *Mann2005* ***McCormick1992*** *Smith2003* *Yalch1976* Zwane2011	***Anderson1988*** *Battaglia1996* *Clausen1968* ***Cohen1985*** Granberg1992 ***Kruse2009*** ***Presser1992*** *Traugott1979* ***Yan2011*** ***Yan2012***		***Rendtel2004***
Direction clear	***Axinn2013*** ***Bailar1975*** ***Bailar1989*** *Bridge1977* Dholakia2002 ***Halpern-Manners2012*** Levav2006 Nancarrow2007 ***Shack-Marquez1986*** Sharpe1998 Sutton1994 ***Torche2012***	***Biderman1984*** Buck1977 Chandon2004 ***Clinton2001*** Coen2005 *Coombs1973* ***Corder1989*** Crespi1948 ***Das2011*** ***Dennis2001*** ***Fendrich1994*** Godin2011 ***Johnston1997*** ***Kruse2009*** ***Lauritsen1998*** ***Mathiowetz1994*** ***Mensch1988*** ***Meurs1989*** Morwitz1993 Parfitt1986 *Percy2005* ***Porst1987*** ***Sturgis2009*** ***Thornberry1989*** ***Toepoel2009*** Toh2006 *Voogt2002* ***Wagstaff2009*** *Wang2000* Williams2006 ***Wilson2005*** ***Woltman1984*** ***Yan2010*** ***Yan2011*** ***Zajacova2010***	***Uhrig2012*** ***Waterton1989***	

Table F2. Continued

	Disadvantageous, Manipulation	Disadvantageous, Correlational	Advantageous, Manipulation	Advantageous, Correlational
Direction unclear	***Ghangurde1982*** ***McCormick1992*** ***Neter1964*** ***Pennell1992*** ***Silberstein1989*** ***Veroff1992*** ***Warren2012*** ***Waterton1989*** Wilson1984 Wilson1993 Zwane2011	***Bartels1999*** ***Binswanger2013*** Christensen2013 ***Clinton2001*** Coen2005 Cohen1985 *Coombs1973* ***Dennis2001*** ***Ehrenberg1960*** ***Golob1990*** ***Kruse2009*** ***Menard1993*** ***Nukulkij2007*** ***Pevalin2000*** ***Pineau2005*** ***Porst1987*** Sen1976 ***Sobol1959*** ***Sturgis2009*** ***Toepoel2009*** ***VanderZouwen2001*** *Wang2000* ***Warren2012*** ***Yan2011***		***Frick2004*** ***Landua1993*** ***Rendtel2004***

Note: Italic font indicates population studies; italic bold indicates studies with 3 or more waves.

Bibliography[38]

American Association for Public Opinion Research. (2011). Standard Definitions: Final Dispositions of Case Codes and Outcome Rates for Surveys. 7th edition.: AAPOR.

Ajzen, I. (1991). The theory of planned behavior. *Organizational behavior and human decision processes, 50*, 179-211.

Allison, P. D. (2009). *Fixed effects regression models*. Thousand Oaks, CA: SAGE.

*Anderson, B. A., Silver, B. D., & Abramson, P. R. (1988). The effects of race of the interviewer on measures of electoral participation by blacks in SRC national election studies. Public Opinion Quarterly, 52, 53-83.

Apodaca, R., Lea, S., & Edwards, B. (1998). *The effect of longitudinal burden on survey participation*. Paper presented at the Annual Conference of American Association for Public Opinion Research St. Louis, Missouri. http://www.amstat.org/sections/srms/Proceedings/papers/1998_156.pdf

Auspurg, K., Burton, J., Cullinane, C., Delavande, A., Fumagelli, L., Iacovou, M., . . . Zafar, B. (2013). Understanding Society Innovation Panel Wave 5: Results from Methodological Experiments. In J. Burton (Ed.), *Understanding Society Working Paper Series*. Essex: Institute for Social and Economic Research, University of Essex.

Axinn, W. G., Jennings, E. A., & Couper, M. P. (2013). *Behavioral response to frequent measurement*. Unpublished paper under review. Survey Research Center, University of Michigan. Ann Arbor, MI.

Bailar, B. A. (1975). Effects of Rotation Group Bias on Estimates from Panel Surveys *Journal of the American Statistical Association, 70*, 23-30.

Bailar, B. A. (1989). Information needs, surveys, and measurement errors. In D. Kasprzyk, G. Duncan, G. Kalton & M. P. Singh (Eds.), *Panel surveys* (pp. 1-24). New York: Wiley.

Baker, R., Blumberg, S. J., Brick, J. M., Couper, M. P., Courtright, M., Dennis, J. M., . . . Zahs, D. (2010). AAPOR Report on Online Panels. Lenexa, Kansas: AAPOR.

Bandilla, W., Kaczmirek, L., Blohm, M., & Neubarth, W. (2009). Coverage und Nonresponse-Effekte bei Online-Bevölkerungsumfragen [Coverage and nonresponse effects in online population-surveys]. In N. Jackob, H. Schoen & T. Zerback (Eds.), *Sozialforschung im Internet: Methodologie und Praxis der Online-Befragung* (pp. 129-143). Wiesbaden: VS Verlag.

Bartels, L. (1999). Panel Effects in the American National Election Studies. *Political Analysis, 8*, 1-20.

Bartsch, S. (2012). *"... würden Sie mir dazu Ihre E-Mail-Adresse verraten?": Internetnutzung und Nonresponse beim Aufbau eines Online Access Panels ["...would you tell me your email adress for this?": Internet Use and Nonresponse during the set up process of the Online Access Panel]*. Baden-Baden: Nomos.

[38] References marked with an asterisk indicate studies which appear solely in the literature review on panel conditioning in Appendix F.

Battaglia, M. P., Zell, E. R., & Ching, P. L. Y. H. (1996). Can participating in a panle sample introduce bias into trend estimates? *Proceedings of the Survey Research Methods Section* (pp. 1010-1013). Alexandira, VA: American Statistical Association.

Behr, A., Bellgardt, E., & Rendtel, U. (2005). Extent and Determinants of Panel Attrition in the European Community Household Panel. *European Sociological Review, 21*(5), 489-512.

Bergmann, M. (2011). Stata-ado "ipfweight": IPF-Algorithm to create adjustment survey weights. from http://fmwww.bc.edu/repec/bocode/i/ipfweight.ado

Bethlehem, J. D. (2008). Surveys without questions. In E. DeLeeuw, J. J. Hox & D. A. Dillman (Eds.), *International handbook of survey methodology* (pp. 500-511). New York: Taylor & Francis Group.

*Biderman, A. D., & Cantor, D. (1984). A longitudinal analysis of bounding, respondent conditioning and mobility as sources of panel bias in the National Crime Survey Proceedings of the Survey Research Section of the American Statistical Association (pp. 708-713). Alexandria, VA.

Biemer, P. P. (1988). Measuring Data Quality. In R. M. Groves, P. P. Biemer, L. E. Lyberg, J. T. Massey, W. L. Nicholls & J. Waksberg (Eds.), *Telephone Survey Methodology* (pp. 273-283). New York: Wiley.

Biemer, P. P. (2010). Total survey error: design, implementation, and evaluation. *Public Opinion Quarterly, 74*(5), 817-848.

Biemer, P. P., & Lyberg, L. E. (2003). *Introduction to survey quality*. Hoboken, New Jersey: Wiley.

Binswanger, J., Schunk, D., & Toepoel, V. (2013). Panel conditioning in difficult attitudinal questions. *Public Opinion Quarterly, Advance Access*(September 16, 2013). doi: 10.1093/poq/nft030

Blasius, J. (1998). Zur Messung von Umweltverhalten. In J. Schupp & G. Wagner (Eds.), *Umwelt und empirische Sozial- und Wirtschaftsforschung [The environment and empirical social and economic research]* (pp. 17-27). Berlin: DIW Berlin [German Institute for Economic Research].

Bosnjak, M. (2003). *(Non)response bei Web-Befragungen [Nonresponse in web surveys]*. Aachen: Shaker-Verlag.

Bosnjak, M., Haas, I., Galesic, M., Kaczmirek, L., Bandilla, W., & Couper, M. P. (2013). Sample composition discrepancies in different stages of a probability-based online panel. *Field Methods*. doi: 10.1177/1525822X12472951

Bradburn, N. M. (1978). *Respondent burden.* Paper presented at the Proceedings of the american Statistical Association, Survey Resaearch Methods Section, Alexandria, VA.

Branden, L., Gritz, M. R., & Pergamit, M. R. (1995). The Effect of Interview Length on Attrition in the National Longitudinal Study of Youth: U.S. Department of Labor Bureau of Labor Statistics

Bridge, G. R., Reeder, L. G., Kanouse, D., Kinder, D. R., Nagy, V. T., & Judd, C. M. (1977). Interviewing changes attitudes -- sometimes. *Public Opinion Quarterly, 41*, 56-64.

&

*Buck, S., Fairclough, E. H., Jephcott, J., & Ringer, D. W. C. (1977). Conditioning and bias in consumer panels - some new results. Journal of the Market Research Society, 19, 59-75.

Buelens, B., Van der Laan, J., Schouten, B., Van den Brakel, J., Burger, J., & Klausch, T. (2012). Disentangling mode-specific selection and measurement bias in social surveys (Vol. 11). The Hague/Herleen: Statistics Netherlands.

Burkam, D. T., & Lee, V. E. (1998). Effects of monotone and nonmonotone attrition on parameter estimates in regressionmodels with educational data: Demographic effects on achievement, aspirations, and attitudes. *The Journal of Human Resources, 33*(2), 555-574.

Callegaro, M. (2013a). From mixed-mode to multiple devices: Web surveys, smartphone surveys and apps: has the respondent gone ahead of us in answering surveys? *International Journal of Market Research, 55*(2), 317-320.

Callegaro, M. (2013b). Paradata in web surveys. In F. Kreuter (Ed.), *Improving surveys with paradata: analytic uses of process information* (pp. 261-280). Hoboken, New Jersey: Wiley.

Callegaro, M., Baker, R., Bethlehem, J. D., Göritz, A. S., Krosnick, J. A., & Lavrakas, P. J. (2014). Online panel research: History, concepts, applications and a look at the future. In M. Callegaro, R. Baker, J. D. Bethlehem, A. S. Göritz, J. A. Krosnick & P. J. Lavrakas (Eds.), *Online panel research: A data quality perspective* (pp. 1-22). New York: Wiley.

Callegaro, M., & DiSogra, C. (2008). Computing Response Metrics for Online Panels. *Public Opinion Quarterly, 72*(5), 1008-1031.

Callegaro, M., Villar, A., Yeager, D., & Krosnick, J. A. (2014). A critical review of studies investigating the quality of data obtained with online panels based on probability and nonprobability samples. In M. Callegaro, R. Baker, J. D. Bethlehem, A. S. Göritz, J. A. Krosnick & P. J. Lavrakas (Eds.), *Online panel research: A data quality perspective* (pp. 23-53). New York: Wiley.

Cantor, D. (1989). Substantive implications of longitudinal design features: the national crime survey as a case study. In D. Kasprzyk, G. Duncan, G. Kalton & M. P. Singh (Eds.), *Panel Surveys* (pp. 25-51). New York: Wiley.

Cantor, D. (2008). A Review and Summary of Studies on Panel Conditioning. In S. Menard (Ed.), *Handbook of Longitudinal Research: Design, Measurement, and Analysis* (pp. 123-138). Burlington, MA: Elsevier.

Castiglioni, L., Pforr, K., & Krieger, U. (2008). The Effect of Incentives on Response Rates and Panel Attrition: Results of a Controlled Experiment. *Survey Research Methods, 2*(3), 151-158.

Chang, L., & Krosnick, J. A. (2009). National surveys via RDD telephone interviewing versus the Internet: Comparing sample representativeness and response quality. *Public Opinion Quarterly, 73*(4), 641-678.

*Chandon, P., Morwitz, V. G., & Reinartz, W. J. (2004). The short- and long-term effects of measuring intent to repurchase. Journal of Consumer Research, 31(3), 566-572. doi: 10.1086/425091

&

*Christensen, A. S., & Ladenburg, J. (2013). Does survey experience affect respondents' reported level of satisfaction? Quality & Quantity, 47(5), 2659-2669. doi: 10.1007/s11135-012-9678-3

Church, A. H. (1993). Estimating the effect of incentives on mail survey response rates: A meta-analysis. *Public Opinion Quarterly, 52*, 67-79.

Clausen, A. R. (1968). Response validity - vote report. Public Opinion Quarterly, 32(4), 588-606. doi: 10.1086/267648

Clinton, J. D. (2001). *Panel Bias from Attrition and Conditioning: A Case Study of the Knowledge Networks Panel.* Retrieved from www.knowledgenetworks.com/insights/docs/Panel%20Effects.pdf.

*Coen, T., Lorch, J., & Piekarski, L. (2005). The Effects of Survey Frequency on Panelists' Responses. Retrieved from http://www.surveysampling.com/ssi-media/Corporate/white_papers/The-Effects-of-Survey-Frequency-on-Panelists-Responses.image

Cohen, J. (1992). A power primer. *Psychological Bulletin, 111*, 155-159.

*Cohen, S. B., & Burt, V. L. (1985). Data collection frequency effect in the National Medical Care Expenditure Survey. Journal of Economic and Social Measurement, 13, 125-151.

Coombs, L. C. (1973). Problems of Contamination in Panel Surveys: A Brief Report on an Independent Sample. Studies in Family Planning, 4(10), 257-261.

Corder, L. S., & Horvitz, D. G. (1989). Panel effects in the national medical care utilization and expenditure survey. In D. Kasprzyk, G. Duncan, G. Kalton & M. P. Singh (Eds.), Panel surveys (pp. 304-317). New York: Wiley.

Couper, M. P. (1998). *Measuring survey quality in CASIC environment*. Paper presented at the Joint Statistical Meetings of the American Statistical Association, Dallas, TX.

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

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

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

Couper, M. P., Singer, E., Conrad, F. G., & Groves, R. M. (2008). Risk of Disclosure, Perceptions of Risk, and Concerns about Privacy and Confidentiality as Factors in Survey Participation. *Journal of Official Statistics, 24*(2), 255–275.

*Crespi, L. (1948). The interview effect in polling. Public Opinion Quarterly, 12, 99-111.

Das, M. (2012). Innovation in online data collection for scientific research: the Dutch MESS project *Methodological Innovations Online* (Vol. 7, pp. 7-24).

*Das, M., Toepoel, V., & van Soest, A. (2011). Nonparametric Tests of Panel Conditioning and Attrition Bias in Panel Surveys. Sociological Methods and Research, 40(1), 32-56.

*Dholakia, U. M., & Morwitz, V. G. (2002). The scope and persistence of mere-measurement effects: Evidence from a field study of customer satisfaction measurement. Journal of Consumer Research, 29(2), 159-167. doi: 10.1086/341568

*Dholakia, U. M., Singh, S. S., & Westbrook, R. A. (2010). Understanding the effects of post-service experience surveys on delay and acceleration of customer purchasing behavior: Evidence from the automotive services industry. Journal of Service Research, 13(4), 362-378. doi: 10.1177/1094670510375105

De Keulenaer, F. (2005). Using process data to predict attrition from a panel survey: A case study *Proceedings of the Survey Research Methods Section* (pp. 3844-3851): American Statistical Association.

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

De Leeuw, E. D., & de Heer, W. (2002). Trends in Household Survey Nonresponse: A Longitudinal and International Comparison. In R. M. Groves, D. A. Dillman, J. L. Eltinge & R. J. A. Little (Eds.), *Survey Nonresponse* (pp. 41-54). New York: Wiley.

De Leeuw, E. D., & Hox, J. J. (2004). I am not selling anything: 29 experiments in telephone introductions. *International Journal of Public Opinion Research, 16*(4), 464-473.

De Leeuw, E. D., & Hox, J. J. (2011). Internet surveys as a part of a mixed-mode design. In M. Das, P. Ester & L. Kaczmirek (Eds.), *Social and behavioral research and the Internet* (Vol. 45-76). New York: Routledge.

Dennis, M. (2001). Are Internet Panels Creating Professional Respondents? *Marketing Research, 13*(2), 34-38.

Destatis [German Federal Statistical Office]. (2011). *Wirtschaftsrechnungen: Private Haushalte in der Informationsgesellschaft. Nutzung von Informations- und Kommunikationstechnologien* [Budget surveys: Private households in the information society: usage of information and communication technology.]. Wiesbaden: Statistisches Bundesamt.

Dillman, D. A. (1978). *Mail and telephone surveys: The total design method*. New York: Wiley.

Dillman, D. A., Eltinge, J. L., Groves, R. M., & Little, R. J. A. (2002). Survey nonresponse in design, data collection, and analysis. In R. M. Groves, D. A. Dillman, J. L. Eltinge & R. J. A. Little (Eds.), *Survey nonresponse* (pp. 3-26). New York: Wiley.

Dillman, D. A., Smyth, J. D., & Christian, L. M. (2009). *Internet, mail, and mixed-mode surveys: The tailored design method* Hoboken, New Jersey: Wiley.

DiSogra, C., & Callegaro, M. (2009). *Computing response rates for probability-based web panels*. Paper presented at the Joint Statistical Meetings.

*Ehrenberg, A. S. C. (1960). A study of some potential biases in the operation of a consumer panel. The Royal Statistical Society Series C-Applied Statistics, 9(1), 20-27.

ESOMAR. (2007). ICC/ESOMAR International Code on Market and Social Research. Retrieved from http://www.esomar.org/uploads/public/knowledge-and-standards/codes-and-guidelines/ICCESOMAR_Code_English_.pdf

ESOMAR. (2009). ESOMAR/WAAPOR Guide to Opinion Polls and Published Surveys. Retrieved from http://www.esomar.org/uploads/public/knowledge-and-standards/codes-and-guidelines/WAPOR-ESOMAR_Guidelines.pdf

ESOMAR. (2011). ESOMAR Guideline for Online Research. Retrieved from http://www.esomar.org/knowledge-and-standards/codes-and-guidelines/guideline-for-online-research.php

ESOMAR. (2012). 28 Questions to Help Byers of Online Samples. Retrieved from http://www.esomar.org/uploads/public/knowledge-and-standards/documents/ESOMAR-28-Questions-to-Help-Buyers-of-Online-Samples-September-2012.pdf

ESOMAR. (2014). ESOMAR/GRBN Guideline For Online Sample Quality. Retrieved from http://www.esomar.org/uploads/public/knowledge-and-standards/codes-and-guidelines/ESOMAR-GRBN-draft-Online-Sample-Quality-Guideline-April-2014.pdf

*Fendrich, M., & Vaughn, C. M. (1994). Diminished lifetime substance use over time - an inquiry into differential underreporting. Public Opinion Quarterly, 58(1), 96-123. doi: 10.1086/269410

*Ferber, R. (1953). Observations on a consumer panel operation. Journal of Marketing, 17(3), 246-259.

Ferguson, C. J. (2009). An Effect Size Primer: A Guide for Clinicians and Researchers. *Professional Psychology: Research and Practice, 40*(5), 532–538.

Fitzgerald, J., Gottschalk, P., & Moffitt, R. (1998). An Analysis of Sample Attrition in Panel Data: The Michigan Study of Income Dynamics *The Journal of Human Resources, 33*(2), 251-299.

*Fitzsimons, G. J., & Morwitz, V. G. (1996). The effect of measuring intent on brand-level purchase behavior. Journal of Consumer Research, 23(1), 1-11. doi: 10.1086/209462

Fox, R. J., Crask, M. R., & Kim, J. (1988). Mail survey response rate: A meta-analysis of selected techniques for inducing response. *Public Opinion Quarterly, 52*(4), 467-491.

Frick, J. R., Goebel, J., Schechtman, E., Wagner, G., & Yitzhaki, S. (2004). Using analysis of Gini (ANoGi) for detecting whether two sub-samples represent the same universe: The SOEP experience *IZA Discussion paper series* (Vol. 1049). Bonn: Institute for the Study of Labor (IZA).

Fricker, S., Galesic, M., Tourangeau, R., & Yan, T. (2005). An experimental comparison of web and telephone surveys. *Public Opinion Quarterly, 69*(3), 370-392.

Gabler, S., & Häder, S. (2002). *Telefonstichproben. Methodische Innovationen und Anwendungen in Deutschland [Telephone samples: Methodological innovations and applications in Germany]*. Münster: Waxmann Verlag.

Gabler, S., Häder, S., Lehnhoff, I., & Mardian, E. (2012). Weighting for the unequal inclusion probabilities and nonresponse in dual frame telephone surveys. In S. Häder, M. Häder & M. Kühne (Eds.), *Telephone surveys in Europe: Research and practice* (pp. 147-167). Heidelberg: Springer.

Galesic, M., & Bosnjak, M. (2009). Effects of questionnaire length on participation and indicators of response quality in a web survey. *Public Opinion Quarterly, 73*(2), 349-360.

&

*Ghanghurde, P. D. (1982). Rotation group bias in the LFS estimates Proceedings of the American Statistical Association (pp. 421-426). Alexandria, VA.

*Godin, G., Belanger-Gravel, A., Amireault, S., Vohl, M. C., & Perusse, L. (2011). The effect of mere-measurement of cognitions on physical activity behavior: a randomized controlled trial among overweight and obese individuals. International Journal of Behavioral Nutrition and Physical Activity, 8. doi: 10.1186/1479-5868-8-2

*Godin, G., Sheeran, P., Conner, M., & Germain, M. (2008). Asking questions changes behavior: Mere measurement effects on frequency of blood donation. Health Psychology, 27(2), 179-184. doi: 10.1037/0278-6133.27.2.179

*Golob, T. F. (1990). The dynamics of household travel time expenditures and car ownership decisions. Transportation Research Part a-Policy and Practice, 24(6), 443-463. doi: 10.1016/0191-2607(90)90035-5

Göritz, A. S. (2006). Incentives in web studies: Methodological issues and a review. *International Journal of Internet Science, 1*, 58-70.

Göritz, A. S. (2008). The Lont-term Effect of Material Incentives on Participation in Online Panels. *Field Methods, 20*(3), 211-225.

Goyder, J. (1987). *The silent minority: Nonrespondents on sample surveys*. Cambridge: Polity Press.

*Granberg, D., & Holmberg, S. (1992). The hawthorne effect in election studies— the impact of survey participation on voting. British Journal of Political Science, 22, 240-247.

Grönlund, K., & Strandberg, K. (2014). Online panels and validity: Representativeness and attrition in the Finnish eOpinion panel. In M. Callegaro, R. Baker, J. D. Bethlehem, A. S. Göritz, J. A. Krosnick & P. J. Lavrakas (Eds.), *Online Panel Research: A Data Quality Perspective* (pp. 86-103). West Sussex: Wiley.

Groves, R. M. (1989). *Survey errors and survey costs*. New York: Wiley.

Groves, R. M., Fowler, F., Couper, M. P., Singer, E., & Tourangeau, R. (2004). *Survey Methodology*. New York: Wiley.

Groves, R. M., Cialdini, R. B., & Couper, M. P. (1992). Understanding the decision to participate in a survey. *Public Opinion Quarterly, 56*, 475-495.

Groves, R. M., & Couper, M. P. (1998). *Nonresponse in Household Interview Surveys*. New York: Wiley.

Groves, R. M., & Lyberg, L. E. (2010). Total survey error: past, present, and future. *Public Opinion Quarterly, 74*(5), 849-879.

Groves, R. M., & Peytcheva, E. (2008). The Impact of Nonresponse Rates on Nonresponse Bias. *Public Opinion Quarterly, 72*(2), 167-179.

Groves, R. M., Singer, E., & Corning, A. D. (2000). Leverage-saliency theory of survey participation. *Public Opinion Quarterly, 64*, 299-308.

Halpern-Manners, A., & Warren, J. R. (2012). Panel Conditioning in Longitudinal Studies: Evidence From Labor Force Items in the Current Population Survey. *Demography, 49*(4), 1499-1519. doi: 10.1007/s13524-012-0124-x

&

Hawkes, D., & Plewis, I. (2006). Modelling Non-response in the National Child Development Study. *Journal of the Royal Statistical Society, 169*(3), 479-491.

Headey, B., & Wearing, A. (1992). *Understanding happiness: A theory of subjective well-being*. Melbourne: Longman Cheshire.

Heberlein, T. A., & Baumgartner, R. (1978). Factors affecting response rates to mailed questionnaires: A quantitative analysis of the published literature. *American Sociological Review, 43*(4), 447-462.

Heerwegh, D. (2005). *Web surveys: Explaining and reducing unit nonresponse, item nonresponse and partial nonresponse (Doctoral dissertation)*. Leuven: Katholieke Universiteit Leuven.

Hill, D. H., & Willis, R. J. (2001). Reducing panel attrition: A search for effective policy instruments. *The Journal of Human Resources, 36*(3), 416-438.

Hirano, K., & Imbens, G. W. (2001). Estimation of causal effects using propensity score weighting: An application to data on right heart catheteritation. *Health Services & Outcomes Research Methodology, 2*, 259-278.

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

Hox, J. J., De Leeuw, E. D., Klausch, T., & Zijlmans, E. (2014). *Measurement equivalence in mixed mode surveys*. Paper presented at the GOR The General Online Research Conference, Cologne, Germany.

Hox, J. J., De Leeuw, E. D., & Vorst, H. (1995). Survey participation as reasoned action: a behavioral paradigm for survey nonresponse? *Bulletin de Méthodologie Sociologique, 48*, 52-67.

Initiative D21, & TNS Infratest. (2013). D21 Digital Index: Auf dem Weg in ein digitales Deutschland? [D21 digital index: on the way to digital Germany?]. Berlin: Initiative D21 & TNS Infratest.

Jäckle, A., & Lynn, P. (2008). Respondent incentives in a multi-mode panel survey: Cumulative effects on nonresponse and bias. *Survey Methodology, 34*(1), 105-117.

*Johnston, L. D., & O'Malley, P. (1997). The recanting of earlier reported drug use by young adults. In L. Harrison & A. Hughes (Eds.), The Validity of Self-Reported Drug Use: Improving the Accuracy of Survey Estimates (pp. 59-80). Washington, DC: National Institute on Drug Abuse.

Joinson, A. N., Paine, C. B., Buchanan, T., & Reips, U.-D. (2007). *Privacy, Trust and Self-Disclosure to Web-Based Surveys.* Paper presented at the Proceedings of the Fifth International Conference of the Association for Survey Computing, Southampton.

Kaczmirek, L. (2009). *Human-survey interaction: usability and nonresponse in online surveys*. Cologne: Herbert von Halem Verlag.

Kaczmirek, L. (2013). *A survey of the uses of paradata.* Paper presented at the MESS workshop, The Hague, The Netherlands.

Kaczmirek, L., Baier, C., & Zuell, C. (2010). Wie empfinden Teilnehmer die Fragen in Online-Befragungen? Entwicklung eines Diktionars für die automatische Codierung freier

&

Antworten [How participants feel about questions in online surveys? Development of the dictionary for the automatic coding of open answers]. In M. Welker & C. Wuensch (Eds.), *Die Online-Inhaltsanalyse: Forschungsobjekt Internet [Online content analysis: Internet as a research object]* (pp. 191-223). Cologne: Herbert von Hamel Verlag.

Kaczmirek, L., & Raabe, J. (2010). Datenquellen und Standarduntersuchungen zur Online-Kommunikation [Data sources and standardized stuies on online communication]. In W. Schwieger & K. Beck (Eds.), *Handbuch Online-Kommunikation [Handbook of the Online Communication]* (pp. 518-540). Wiesbaden: VS Verlag.

Kalton, G., & Citro, C. F. (1993). Panel Surveys: Adding the Fourth Dimension. *Survey Methodology, 19*(2), 205-260.

Kalton, G., Kasprzyk, D., & McMillen, D. B. (1989). Nonsampling errors in panel surveys. In D. Kasprzyk, G. Duncan, G. Kalton & M. P. Singh (Eds.), *Panel Surveys* (pp. 249-270). New York: Wiley.

Klausch, T., Hox, J. J., & Schouten, B. (2014). *Evaluating mixed-mode redesign strategies against benchmark surveys: The case of the Crime Victimization Survey*. Paper presented at the GOR General Online Conference, 05-07.03.2014 in Cologne.

Kohut, A., Keeter, S., Doherty, C., Dimock, M., & Christian, L. (2012). *Assessing the representativeness of public opinion surveys.*Washington, DC: The Pew Research Center for the People and the Press.

Kraut, R. E., & McConahay, J. B. (1973). How Being Interviewed Affects Voting: An Experiment. *Public Opinion Quarterly, 37*(3), 398-406.

Kreuter, F. (2013). Improving surveys with paradata: Introduction. In F. Kreuter (Ed.), *Improving surveys with paradata: Analytic uses of process information* (pp. 1-12). Hoboken, New Jersey: Wiley.

Kreuter, F., McCulloch, S., Presser, S., & Tourangeau, R. (2011). The effects of asking filter questions in interleafed versus grouped format. *Sociological Methods and Research, 40*, 88-104. doi: 10.1177/0049124110392342

Kreuter, F., & Olson, K. (2013). Paradata for nonresponse error investigation. In F. Kreuter (Ed.), *Improving surveys with paradata: Analytic uses of process information* (pp. 13-42). Hoboken, New Jersey: Wiley.

Kreuter, F., Presser, S., & Tourangeau, R. (2008). Social desirability bias in CATI, IVR, and Web surveys: The effects of mode and question sensitivity. *Public Opinion Quarterly, 72*, 847-865. doi: 10.1093/poq/nfn063

Krosnick, J. A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology, 5*, 213-236.

Krosnick, J. A., & Alwin, D. F. (1987). An evaluation of a cognitive theory of response order effects in survey measurement *Public Opinion Quarterly, 51*, 201-219.

Kruse, Y., Callegaro, M., Dennis, M. J., DiSogra, C., Subias, S., Lawrence, M., & Tompson, T. (2009). *Panel Conditioning and Attrition in the AP-Yahoo! News Election Panel Study*

&

Retrieved from http://www.knowledgenetworks.com/ganp/docs/jsm2009/Panel%20Conditioning%20and%20Attrition_JSM_2009_submitted.pdf

*Landua, D. (1993). Changes in replies to questions about satisfaction in panel surveys - an analysis of nonintentional effects of the longitudinal design. Kolner Zeitschrift Fur Soziologie Und Sozialpsychologie, 45(3), 553-571.

Laurie, H. (2007). The effect of increasing financial incentives in a panel survey: An experiment on the British Household Panel Survey, Wave 14 *ISER Working Paper* (Vol. 2007-05). Colchester: University of Essex.

Laurie, H., & Lynn, P. (2009). The Use of Respondent Incentives on Longitudinal Surveys. In P. Lynn (Ed.), *Methodology of longitudinal surveys* (pp. 205-233). West Sussex: Wiley.

Laurie, H., Smith, R., & Scott, L. (1999). Strategies for Reducing Nonresponse in a Longitudinal Panel Survey. *Journal of Official Statistics, 15*(2).

*Lauritsen, J. (1998). The Age-Crime Debate: Assessing the Limits of Longitudinal Self-Report Data. Social Forces, 77(1), 127-155.

Lazarsfeld, P. F. (1940). "Panel" studies. *Public Opinion Quarterly, 4*, 122-128.

Lee, E., Hu, M. Y., & Toh, R. S. (2004). Respondent non-cooperation in surveys and diaries: an analysis of item non-response and panel attrition. *International Journal of Market Research, 46*(3), 311-326.

Lee, S. (2006). An Evaluation of Nonresponse and Coverage Errors in a Prerecruited Probability Web Panel Survey. *Social Science Computer Review, 24*, 460-475.

*Lefevere, E., Huysentruyt, M., Theeten, H., & Van Damme, P. (2010). The effect of mere measurement on vaccination behaviour: A randomised experiment in Flanders (Belgium). Psychology & Health, 25, 53-53.

Lengacher, J. E., Sullivan, C. M., Couper, M. P., & Groves, R. M. (1995). *Once reluctant, always reluctant? Effects of differential incentives on later survey participation in a longitudinal study*. Paper presented at the Annual Conference of the American Association for Public Opinion Research, Fort Lauderdale, Florida.

Lenzner, T., Kaczmirek, L., & Galesic, M. (2011). Seeing through the eyes of the respondent: An eye-tracking study on survey question comprehension. *International Journal of Public Opinion Research, 23*, 361-373.

Lemay, M. (2009). *Understanding the mechanism of panel attrition.* (Doctoral dissertation), University of Maryland. Retrieved from http://drum.lib.umd.edu/bitstream/1903/9631/1/Lemay_umd_0117E_10679.pdf

Lepkowski, J. M., & Couper, M. P. (2002). Nonresponse in the Second Wave of Longitudinal Household Surveys. In R. M. Groves, D. A. Dillman, J. L. Etlinge & R. J. A. Little (Eds.), *Survey Nonresponse* New York: Wiley.

*Levav, J., & Fitzsimons, G. J. (2006). When Questions Change Behavior—The Role of Ease of Representation. Psychological Science, 17, 207-213.

Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis*. Thousand Oaks, California: Sage.

Lozar Manfreda, K., Bosnjak, M., Berzelak, J., Haas, I., & Vehovar, V. (2008). Web surveys versus other survey modes: A meta-analysis comparing response rates. *International Journal of Market Research, 50*(1), 79-104.

Loosveldt, G., Pickery, J., & Billiet, J. (2002). Item nonresponse as a predictor of unit nonresponse in a panel survey. *Journal of Official Statistics, 18*(4), 545-557.

Lugtig, P. (2012). *I think I know what you did last summer: Improving data quality in panel surveys.* (Doctoral Dissertation), Utrecht University.

Lugtig, P. (2014). Panel attrition: separating stayers, fast attriters, gradual attriters and lurkers. *Sociological Methods & Research, Advance online publication*. doi: 10.1177/0049124113520305

Lugtig, P., Das, M., & Scherpenzeel, A. (2014). Nonresponse and attrition in a probability-based online panel for the general population. In M. Callegaro, R. Baker, J. D. Bethlehem, A. S. Göritz, J. A. Krosnick & P. J. Lavrakas (Eds.), *Online Panel Research: A Data Quality Perspective* (pp. 135-153). West Sussex: Wiley.

Lynn, P. (2008). The problem of nonresponse. In E. D. de Leeuw, J. J. Hox & D. A. Dillman (Eds.), *International Handbook of Survey Methodology* (pp. 35-55). East Sussex: Taylor & Francis Group.

Lynn, P. (2009). Methods for longitudinal surveys. In P. Lynn (Ed.), *Methodology of longitudinal surveys* (pp. 1-19). West Sussex: Wiley.

Lynn, P., Buck, N., Burton, J., Jäckle, A., & Laurie, H. (2005). A review of methodological research pertinent to longitudinal survey design and data collection *Working Papers of the Institute for Social and Economic Research* (Vol. 2005-29). Colchester: University of Essex.

Lynn, P., & Clarke, P. (2002). Separating Refusal Bias and Non-Contact Bias: Evidence from UK National Surveys. *Journal of the Royal Statistical Society, Series D, 51*(3), 319-333.

*Mann, C. B. (2005). Unintentional voter mobilization: Does participation in preelection surveys increase voter turnout? Annals of the American Academy of Political and Social Science, 601, 155-168. doi: 10.1177/0002716205278151

Mathiowetz, N. A., & Liar, T. J. (1994). Getting better? Changes or errors in the measurement of functional limitations. *Journal of Economic and Social Measurement, 20*, 237-262.

Matzat, U., & Snijders, C. (2010). Does the online collection of ego-centered network data reduce data quality? An experimental comparison. *Social Networks, 32*, 105-111.

*McCormick, M., Butler, D., & Singh, R. (1992). Investigating time in sample effect for the survey of income and program participation Proceedings of the Survey Research Methods Section of the American Statistical Association. Alexandria, VA.

*Menard, S., & Elliott, D. S. (1993). Data set comparability and short-term trends in crime and delinquency. Journal of Criminal Justice, 21(5), 433-445. doi: 10.1016/0047-2352(93)90030-q

*Mensch, B. S., & Kandel, D. B. (1988). Underreporting of substance use in a national longitudinal youth cohort - individual and interviewer effects. Public Opinion Quarterly, 52(1), 100-124. doi: 10.1086/269084

&

*Meurs, H., Van Wissen, L., & Visser, J. (1989). Measurement biases in panel data. Transportation, 16, 175-194.

Mohorko, A., De Leeuw, E. D., & 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.

*Morrison, D. G., Frank, R. E., & Massy, W. F. (1966). Note on panel bias. Journal of Marketing Research, 3(1), 85-88. doi: 10.2307/3149440

*Morwitz, V. G., Johnson, E., & Schmittlein, D. (1993). Does measuring intent change behavior. Journal of Consumer Research, 20(1), 46-61. doi: 10.1086/209332

Nancarrow, C., & Cartwright, T. (2007). Online access panels and tracking research: The conditioning issue. *International Journal of Market Research, 49*, 573-594.

*Neter, J., & Waksberg, J. (1964). Conditioning effects from repeated household interviews. Journal of Marketing, 28(2), 51-56. doi: 10.2307/1248891

Neugebauer, B. (2004). Die Erfasung von Umweltbewusstsein und Umveltverhalten [Measurement of environmental awareness and ecological behavior]. Mannheim, Germany: ZUMA.

Nicolaas, G., Calderwood, L., Lynn, P., & Roberts, C. (2014). Web surveys for the general population: How, why and when? London: National Centre for Research Methods

Nicoletti, C., & Peracchi, F. (2005). Survey response and survey characteristics: microlevel evidence from the European Community Household Panel. *Journal of the Royal Statistical Society Series A, 168*, 763-781.

NTIA - National Telecommuniation and Information Administration, & Economics and Statistics Administration. (2013). Exploring the digital nation: America's emerging online experience. Washington, DC: US Department of Commerce.

*Nukulkij, P., Hadfield, J., Subias, S., & Lewis, S. (2007). An investigation of panel conditioning with attitudes toward U.S. foreign policy. Paper presented at the Annual Conference of the American Association for Public Opinion Research, Anaheim, California.

*O'Sullivan, I., Orbell, S., Rakow, T., & Parker, R. (2004). Prospective Research in Health Service Settings: Health Psychology, Science and the 'Hawthorne' Effect. Journal of Health Psychology, 9(3), 355-359.

Olsen, R. J. (2005). The problem of respondent attrition: Survey methodology is key. *Monthly Labor Review, 128*, 63-70.

Olson, K., & Witt, L. (2011). Are we keepinh the people who used to stay? Changes in correlates of panel survey attrition over time. *Social Science Research, 40*(4), 1037-1050.

*Parfitt, J. (1986). Panel research. In R. M. Worcester & J. Downham (Eds.), Consumer Market Research Handbook: Elsevier.

Pennell, S. G., & Lepkowski, J. M. (1992). Panel conditioning effects in the Survey of Income and Program Participation *Proceedings of the Survey Research Methods Section* (pp. 566-571). Alexandria, VA: American Statistical Association.

Peracchi, F. (2002). The European Community Household Panel: A review. *Empirical Economics, 27*, 63-90.

&

*Percy, A., McAlister, S., Higgins, K., McCrystal, P., & Thornton, M. (2005). Response consistency in young adolescents' drug use self-reports: A recanting rate analysis. Addiction, 100, 189–196.

*Pevalin, D. (2000). Multiple applications of the GHQ-12 in a general population sample: an investigation of long-term retest effects. Social psychiatry and psychiatric epidemiology, 35, 508-512.

Peytchev, A. (2009). Survey breakoff. *Public Opinion Quarterly, 73*, 74-97.

*Pineau, V., Nukulkij, P., & Tang, X. (2005). Assessing Panel Bias in the Knowledge Networks Panel: Updated Results from 2005 Research. Paper presented at the Proceedings of the Section on Survey Research Methods of the American Statistical Association, Alexandria, VA.

*Porst, R., & Zeifang, K. (1987). A Description of the German General Social Survey Test-Retest Study and a Report on the Stabilities of the Sociodemographic Variables. Sociological Methods & Research, 15, 177-218.

*Prais, S. J. (1958). Some problems in the measurement of price changes with special reference to the cost of living - discussion. Journal of the Royal Statistical Society Series a-General, 121(3), 312-332. doi: 10.2307/2342903

*Presser, S., & Traugott, M. (1992). Little white lies and social science models: Correlated response errors in a panel study of voting. Public Opinion Quarterly, 56, 77-86.

Rabe-Hesketh, S., & Skrondal, A. (2012). *Multilevel and longitudinal modeling using Stata Vloume II: Categorical responses, counts, and survival*. College Station, TX: Stata Press.

Rendtel, U., Nordberg, L., Jäntti, M., Hanisch, J., & Basic, E. (2004). Report on quality of income data. https://www.destatis.de/DE/Methoden/Methodenpapiere/Chintex/ResearchResults/Downloads/WorkingPaper21.pdf?__blob=publicationFile

Rodgers, W. (2011). Effects of increasing the incentive size in a longitudinal study. *Journal of Official Statistics, 27*(2), 279-299.

Rosenbaum, P. R. (1987). Model-based direct adjustment. *Journal of the American Statistical Association, 82*, 387-394.

Rosenbaum, P. R. (2010). *Design of observational studies*. New York: Springer.

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika, 70*, 41-55.

Rubin, D. B. (1978). *Multiple imputations in sample surveys: A phenomenological bayesian approach to nonresponse.* Paper presented at the Procedings of the Survey Research Methods Section, Alexandria, VA.

Rubin, D. B. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association, 91*(434), 473-489.

Sakshaug, J., & Kreuter, F. (2011). Using paradata and other auxiliary data to examine mode switch nonresponse in a "recruit-and-switch" telephone survey. *Journal of Official Statistics, 27*(2), 339–357.

&

*Sandberg, T., & Conner, M. (2009). A mere measurement effect for anticipated regret: Impacts on cervical screening attendance. British Journal of Social Psychology, 48(2), 221-236. doi: 10.1348/014466608x347001

Saris, W. E. (1998). Ten Years of Interviewing without Interviewers: The Telepanel. In M. P. Couper, R. Baker, J. D. Bethlehem, C. Z. F. Clark, J. Martin, I. Nicholls, William L. & J. M. O'Reilly (Eds.), *Computer assisted survey information collection* (pp. 409-429). New York: Wiley.

Schafer, J. L., & Kang, J. (2008). Average causal effects from nonrandomized studies: a practical guide and simulated example. *Psychological Methods, 13*, 279-313.

Schaurer, I., Struminskaya, B., & Kaczmirek, L. (2014). Incentives on demand in a probability-based online panel: Redemption and the choice between payout options. Poster presented at the GOR General Online Conference, 05-07.03.2014 in Cologne.

Scherpenzeel, A., & Bethlehem, J. D. (2011). How representative are online panels? Problems of coverage and selection and possible solutions. In M. Das, P. Ester & L. Kaczmirek (Eds.), *Social and Behavioral Research and the Internet* (pp. 105-130). New York: Routledge.

Scherpenzeel, A., & Das, M. (2011). "True" longitudinal and probability-based Internet panels: Evidence from the Netherlands. In M. Das, P. Ester & L. Kaczmirek (Eds.), *Social and behavioral research and the Internet* (pp. 77-104). New York: Routledge.

Scherpenzeel, A., & Toepoel, V. (2012). Recruiting a probability sample for an online panel: Effects of contact mode, incentives, and information. *Public Opinion Quarterly*.

Scherpenzeel, A., & Toepoel, V. (2014). Informing panel members about study results: Effects of traditional and innovative forms of feedback on participation. In M. Callegaro, R. Baker, J. D. Bethlehem, A. S. Göritz, J. A. Krosnick & P. J. Lavrakas (Eds.), *Online panel research: A data quality perspective* (pp. 193-213). West Sussex: Wiley.

Scherpenzeel, A., & Zandvliet, R. (2011). Slapers en inactieven binnen online panels [Sleepers and inactive participants in online panels] *Ontwikkelingen in het marktonderzoek: Jaarboek MarktOnderzoekAssociatie* (Vol. 36, pp. 189-204). Hout Haarlem: Spaar.

Schnell, R. (1997). *Nonresponse in Bevölkerungsumfragen: Ausmaß, Entwicklung und Ursachen [Nonresponse in general population surveys: magnitude, evolution and reasons]*. Opladen: Leske + Budrich.

Schonlau, M. (2012). Stata-ado package "hotdeckvar" - Single hotdeck imputation. from http://www.schonlau.net/stata/

Schonlau, M., Van Soest, A., & Kapteyn, A. (2007). Are "webographic" or attitudinal questions useful for adjusting estimates from web surveys using propensity scoring? *Survey Research Methods, 1*, 155-163.

Schonlau, M., Van Soest, A., Kapteyn, A., & Couper, M. P. (2009). Selection bias in web surveys and the use of propensity scores *Sociological Methods & Research, 37*, 291-318.

*Shackmarquez, J. (1986). Effects of repeated interviewing on estimation of labor-force status. Journal of Economic and Social Measurement, 14(4), 379-398.

&

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton Mifflin Company.

Sharp, L. M., & Frankel, J. (1983). Respondent burden: A test of some common assumptions. *Public Opinion Quarterly, 47*, 36-53.

*Sharpe, J. P., & Gilbert, D. G. (1998). Effects of repeated administration of the Beck Depression Inventory and other measures of negative mood states. Personality and Individual Differences, 24(4), 457-463. doi: 10.1016/s0191-8869(97)00193-1

Short, K., & McArthur, E. (1986). Life events and sample attrition in the Survey of Income and Program Participation *Proceedings of the Section on Survey Research Methods*. Chicago, IL: American Statisticl Association.

Sikkel, D., & Hoogendoorn, A. (2008). Panel surveys. In E. D. De Leeuw, J. J. Hox & D. A. Dillman (Eds.), *International Handbook of Survey Methodology* (pp. 479-499). New York: Taylor & Francis Group.

Silberstein, A. R., & Jacobs, C. A. (1989). Symptoms of repeated interview effects in the consumer expenditure interview surveys. In D. Kasprzyk, G. Duncan, G. Kalton & M. P. Singh (Eds.), *Panel Surveys* (pp. 289-303). New York: Wiley.

Singer, E. (2002). The use of incentives to reduce nonresponse in household surveys. In R. M. Groves, D. A. Dillman, J. L. Eltinge & R. J. A. Little (Eds.), *Survey nonresponse* (pp. 163-176). New York: Wiley.

Singer, E. (2011). Toward a benefit-cost theory of survey participation: evidence, further tests, and implications. *Journal of Official Statistics, 27*(2), 379-392.

Singer, E., Groves, R. M., & Corning, A. D. (1999). Differential incentives: Beliefs about practices, perceptions of equity, and effects on survey participation. *Public Opinion Quarterly, 63*(2), 251-260.

Singer, E., & Presser, S. (2008). Privacy, confidentiality, and respondent burden as factors in telephone survey nonresponse. In J. M. Lepkowski, C. Tucker, J. M. Brick, E. D. D. Leeuw, L. Japec, P. J. Lavrakas, M. Link & R. L. Sangster (Eds.), *Advances in telephone survey methodology* (pp. 449-470). Hoboken, New Jersey: Wiley.

Singer, E., Van Hoewyk, J., Gebler, N., Raghunathan, T. E., & McGonagle, K. (1999). The effect of incentives on response rates in interviewer-mediated surveys. *Journal of Offcial Statistics, 15*, 217-230.

Singer, E., & Ye, C. (2013). The use and effects of incentives in surveys. *The ANNALS of the American Academy of Political and Social Science, 645*, 112-141.

*Smith, J. P., Gerber, A. S., & Orlich, A. (2003). Self-prophecy effects and voter turnout: An experimental replication. Ploitical Psychology, 24(593-604).

Smith, T. (2003). An experimental comparison of Knowledge Networks and the GSS. *International Journal of Public Opinion Research, 15*, 167-179.

Smith, T. W. (2011). Refining the total survey error perspective. *International Journal of Public Opinion Research, 23*, 464-484.

&

*Sobol, M. G. (1959). Panel mortality and panel bias. Journal of the American Statistical Association, 54(285), 52-68. doi: 10.2307/2282139

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

Stoop, I. A. L. (2005). *The hunt for the last respondent*. The Hague: Social and Cultural Planning Office of the Netherlands.

Struminskaya, B., Kaczmirek, L., Schaurer, I., & Bandilla, W. (2014). Assessing representativeness of a probability-based online panel in Germany. In M. Callegaro, P. J. Lavrakas, J. A. Krosnick, R. Baker, J. D. Bethlehem & A. S. Göritz (Eds.), *Online panel research: A data quality perspective* (pp. 61-85). New York: Wiley.

Stull, G. (2011, 01.04.2011). Nuclear safety in the EU. Library Briefing for the European Parliament. from http://www.europarl.europa.eu/RegData/bibliotheque/briefing/2011/110159/LDM_BRI(2011)110159_REV2_EN.pdf

Sturgis, P., Allum, N., & Brunton-Smith, I. (2009). Attitudes Over Time: The Psychology of Panel Conditioning. In P. Lynn (Ed.), *Methodology of Longitudinal Surveys* (pp. 113-126). New York: Wiley.

*Sutton, S., Bickler, G., Sancho-Aldridge, J., & Saidi, J. (1994). Prospective Study of Predictors of Attendance for Breast Screening in Inner London Journal of Epidemiology and Community Health, 48, 65-73.

*Thornberry, T. P. (1989). Panel effects and the use of self-reported measures of delinquency in longitudinal studies. In M. W. Klein (Ed.), Cross-National Research in Self-Reported Crime and Delinquency (pp. 347-369): Springer Netherlands.

Toepoel, V., Das, M., & van Soest, A. (2009). Relating Question Type to Panel Conditioning: Comparing Trained and Fresh Respondents. *Survey Research Methods, 3*(2), 73-80.

*Toh, R. S., Lee, E., & Hu, M. Y. (2006). Social desirability bias in diary panels is evident in panelists' behavioral frequency. Psychological Reports, 99(2), 322-334. doi: 10.2466/pr0.99.2.322-334

Torche, F., Warren, J. R., Halpern-Manners, A., & Valenzuela, E. (2012). Panel Conditioning in a Longitudinal Study of Adolescents' Substance Use: Evidence from an Experiment. *Social Forces, 90*(3), 891-918. doi: 10.1093/sf/sor006

Tortora, R. D. (2009). Attrition in Consumer Panels. In P. Lynn (Ed.), *Methodology of Longitudinal Surveys* (pp. 235–249). New York: Wiley.

Tourangeau, R. (2004). Survey research and societal change. *Annual Review of Psychology, 55*, 775-801.

Tourangeau, R., Conrad, F. G., & Couper, M. P. (2013). *The science of web surveys*. Oxford: Oxford University Press.

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

&

Trappmann, M., Gramlich, T., & Mosthaf, A. (2012). *The effect of life events between waves on panel attrition*. Paper presented at the 67th Annual Conference of the American Association for Public Opinion Research, Orlando.

Traugott, M. W., & Katosh, J. P. (1979). Response validity in surveys of voting behavior. *Public Opinion Quarterly, 43*, 359-377.

Trussell, N., & Lavrakas, P. J. (2004). The influence of incremental increase in token cash incentives on mail survey response: is there an optimal amount? *Public Opinion Quarterly, 68*(3), 349-367.

Uhrig, S. N. (2008). The nature and causes of attrition in the British Household Panel Survey. Essex: Institute for Social and Economic Research University of Essex.

Uhrig, S. N. (2012). Understanding panel conditioning: an examination of social desirability bias in self-reported height and weight in panel surveys using experimental data. *Longitudinal and Life Course Studies, 3*(1), 120-136.

Van den Berg, G. J., Lindeboom, M., & Dolton, P. J. (2006). Survey non-response and the duration of unemployment. *Journal of Royal Statistical Society A, 169*, 585-604.

Van der Zouwen, J., & Van Tilburg, T. (2001). Reactivity in panel studies and its concequences for testing causal hypotheses. *Sociological Methods and Research, 30*, 35-56.

*Van Dongen, A., Abraham, C., Ruiter, R. A. C., & Veldhuizen, I. J. T. (2013). Does questionnaire distribution promote blood donation? An investigation of question-behavior effects. Annals of Behavioral Medicine, 45(2), 163-172. doi: 10.1007/s12160-012-9449-3

Veenhoven, R. (1996). Developments in satisfaction research. *Social Indicators Research, 37*, 1-46.

Vehovar, V., Lozar-Manfreda, K., Zaletel, M., Batagelj, Z. . (2002) . Nonresponse in Web Surveys . In Groves, R.M., Dillman, D. A., Eltinge, J.L., Little, R.J.A. (Eds.), *Survey Nonresponse* (pp. 229-242). New York: Wiley.

Vehre, H. (2011). *"Sie wollen mir doch was verkaufen!": Analyse der Umfrageteilnahme in einem offline rekrutierten Access Panel ["You do want to sell me something!": The analysis of survey participation in an offline-recruited access panel]*. Wiesbaden: VS Verlag.

*Veroff, J., Hatchett, S., & Elizabeth, D. (1992). Consequences of participating in a longitudinal study of marriage. Public Opinion Quarterly, 56, 315-327.

*Voogt, R. J. J., & Van Kempen, H. (2002). Nonresponse bias and stimulus effects in the Dutch National Election Study. Quality & Quantity, 36(4), 325-345.

Wagner, G. G., Pischner, R., & Haisken-DeNew, J. P. (2005). The changing digital divide in Germany. In B. Wellman & C. Haythornthwaite (Eds.), *The Internet in everyday life* (pp. 164-185). Malden, MA: Blackwell Publishing.

Walston, J. T., Lissitz, R. W., & Rudner, L. M. (2006). The influence of web-based questionnaire presentation variations on survey cooperation and perceptions of survey quality. *Journal of Official Statistics, 22*(2), 271-291.

Warren, J. R., & Halpern-Manners, A. (2012). Panel Conditioning Effects in Longitudinal Social Science Surveys. *Sociological Methods and Research, 41*(4), 491-534.

&

Wasmer, M., Scholz, E., Blohm, M., Walter, J., & Jutz, R. (2012). Konzeption und Durchführung der "Allgemeinen Bevölkerungsumfrage der Sozialwissenschaften" (ALLBUS) 2010. [Concept and implementation of the German General Social Survey (ALLBUS) 2010] *GESIS Technical Report* (Vol. 2012/12). Mannheim: GESIS – leibniz-Institute for the Social Sciences.

Watson, D. (2003). Sample attrition between waves 1 and 5 in the European Community Household Panel. *European Sociological Review, 19*(4), 361-378.

Watson, N., & Wooden, M. (2009). Identifying factors affecting longitudinal survey response In P. Lynn (Ed.), *Methodology of Longitudinal Surveys* (pp. 157-181). West Sussex: Wiley.

Watson, N., & Wooden, M. (2011). Re-engaging with survey non-respondents: The BHPS, SOEP and HILDA survey experience *SOEP Papers* (Vol. 379). Berlin: DIW Berlin.

Waterton, J., & Livesley, D. (1989). Evidence of conditioning effects in the British Social Attitudes Panel. In D. Kasprzyk, G. Duncan, G. Kalton & M. P. Singh (Eds.), *Panel Surveys* (pp. 319-339). New York: Wiley.

*Williams, P., Block, L. G., & Fitzsimons, G. J. (2006). Simply asking questions about health behaviors increases both healthy and unhealthy behaviors. Social Influence, 1(2), 117-127.

*Wilson, S. E., & Howell, B. L. (2005). Do panel surveys make people sick? US arthritis trends in the health and retirement study. Social Science & Medicine, 60, 2623-2627.

*Wilson, T. D., Dunn, D. S., Bybee, J. A., Hyman, D. B., & Rotondo, J. A. (1984). Effects of analyzing reasons on attitude-behavior consistency. Journal of Personality and Social Psychology, 47(1), 5-16. doi: 10.1037//0022-3514.47.1.5

*Woltman, H., & Bushery, J. (1984). Summary of results from the National Crime Survey panel bias study. In R. G. Lehnen & W. G. Skogan (Eds.), The National Crime Survey Working Papers (Vol. Volume 11: Methodological Studies, pp. 98-101). Washington, DC: US Department of Justice.

World Nuclear Association. (2011a, 2011). Archive 2011 Germany. Retrieved 15.08.2013, 2013, from http://world-nuclear.org/WNA/Publications/Weekly-Digest/Archive/Archive-2011/#.Ug44qXeP2ik

World Nuclear Association. (2011b, July 2011). World nuclear power reactors & uranium requirements archive. Retrieved 15.08., 2013, from http://world-nuclear.org/info/Facts-and-Figures/World-Nuclear-Power-Reactors-Archive/Reactor-Archive-July-2011/#.Ug4wI3eP2ik

World Nuclear Association. (2012). Waste management: overview. from http://world-nuclear.org/info/Nuclear-Fuel-Cycle/Nuclear-Wastes/Waste-Management-Overview/#.Ug417XeP2il

Yalch, R. (1976). Pre-Election Interview Effects on Voter Turnout *Public Opinion Quarterly, 40*(3), 331-336.

Yan, T., & Copeland, K. (2010). Panel conditioning in the consumer expenditure quarterly interview survey *Proceedings of the Survey Research Methods Section* (pp. 3618-3628). Alexandria, VA: American Statistical Association.

&

Yan, T., Datta, R., & Hepburn, P. (2011). *Conditioning Effects in Panel Participation*. Paper presented at the 66th Annual Conference of the American Association for Public Opinion Research, Phoenix, AZ.

Yan, T., & Eckman, S. (2012). Panel conditioning: Change in true value versus change in self-report *Proceedings of the Survey Research Methods Section* (pp. 4726-4736). Alexandria, VA: American Statistical Association.

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

Zabel, J. E. (1998). An Analysis of Attrition in the Panel Study of Income Dynamics and the Survey of Income and Program Participation with an Application to a Model of Labor Market Behavior. *The Journal of Human Resources, 33*(2).

Zagorsky, J. L., & Rhoton, P. (2008). The effects on promised monetary incentives on attrition in a long-term panel survey. *Public Opinion Quarterly, 72*(3), 502-513.

*Zajacova, A., Dowd, J. B., Schoeni, R. F., & Wallace, R. B. (2010). Consistency and precision of cancer reporting in a multiwave national panel survey. Population Health Metrics, 8. doi: 10.1186/1478-7954-8-20

*Zwane, A. P., Zinman, J., Van Dusen, E., Pariente, W., Null, C., Miguel, E., . . . Banerjee, A. (2011). Being surveyed can change later behavior and related parameter estimates. Proceedings of the National Academy of Sciences of the United States of America, 108(5), 1821-1826. doi: 10.1073/pnas.1000776108

English summary

A survey is a research method for collecting information from a sample of entities, usually persons, in a systematic manner. Information collected in surveys is used for a variety of purposes. Researchers, such as sociologists, political scientists, and economists use this information to study social phenomena and societal processes. Politicians use the results of surveys to guide their decisions in policy-making. Commercial companies use the information gathered in surveys to measure customer preferences or customer satisfaction in order to improve the existing or to develop new products and services. A special type of surveys is the pre-election poll that informs the candidates and the public about political issues and voting intentions.

The information collected in surveys needs to have high quality in order for organizations and individuals who use this information (data users) to make correct decisions. This primarily means that the organizations conducting surveys (producers of data) need to make sure the data is free of errors. On the other hand, the producers of data have limited budgets and resources, thus, they need to maximize the quality under these constraints.

Surveys administered over the Internet offer cost reductions when these surveys are compared to the surveys that use more traditional modes of data collection (face-to-face, telephone, and mail surveys). Face-to-face and telephone surveys are administered by interviewers, so Internet surveys offer an advantage of not having interviewer payments. Mail surveys involve printing and postage expenses as well as the costs of processing the returned questionnaires, whereas in online surveys data provided by respondents are directly available in electronic format. As Internet access rates in many countries are rising rapidly, online surveys and especially online panel surveys (surveys in which persons are asked to complete regularly over specified time intervals) are a lucrative alternative for surveys carried out in traditional modes.

However, some characteristics of online panel surveys (online panels) may cause various errors, threatening the data quality. The main question that this dissertation seeks to answer is that of how good the data collected in online panels is. There are two main types of online panels: those which allow everyone to join (volunteer, nonprobability online panels) and those where the respondents are selected by the researchers as a result of application of statistical procedures of random sampling (probability-based online panels). The latter type allows for drawing valid conclusions about the population of interest. This dissertation concentrates only on probability-based online panels.

Several steps are needed to recruit individuals into a probability-based online panel. First, the researcher has to define the population he or she wants to study and to find a sampling frame. For face-to-face or mail surveys a population register can be such a frame, from which study participants can be randomly drawn. However, these registers (if they are available) do not contain email addresses. Therefore, sampled study participants need to be contacted in an alternative mode (face-to-face or via mail)A similar procedure can be implemented by telephone, using Random Digit Dialing. In some countries, address-based sampling is an alternative to population registers of individuals. Naturally, not

&

everybody in the population has Internet access or uses it. If a researcher wants to study the whole population and not Internet users only, excluding the non-Internet population may result in coverage error. Specifically, coverage error will arise if Internet users differ from non-users in the aspects important for the researcher. Non-users may be offered an alternative completion mode or they may be provided with the equipment necessary for online participation (e.g., a simple computer and Internet access).

Additional problems arise when persons selected for the study cannot be reached (noncontact) of if they could be reached, cannot or do not want to participate in the study (nonresponse). If those persons differ systematically from study participants, it gives rise to a nonresponse error. In panel surveys, nonresponse is even more critical, since study participants who agreed to participate initially may chose not to do so in later waves (attrition). Furthermore, by taking surveys regularly, respondents may learn how to manipulate the questionnaire (for example, answer dishonestly or answer negatively to filter questions) to reduce the burden of participation. This will negatively affect the measurements (measurement error). All errors mentioned above can accumulate in the survey estimates (the final product of a survey) and the conclusions based on these estimates may therefore be misleading. The errors described above are not unique to online panels. They also apply to cross-sectional (one time surveys) or panel surveys carried out in other modes. However, there are some specific problems with online surveys: for example, when a selected respondent refuses to participate because of insufficient skills or experience with the Internet.

The five studies that comprise this dissertation all address the questions of data quality. Each chapter is a single study. Some chapters concentrate on the quality of estimates, whereas other chapters study in more detail the processes which lead to errors in estimates. Data from a probability-based telephone-recruited online panel of Internet users in Germany (GESIS Online Panel Pilot) is used in all the studies. The GESIS Online Panel Pilot was recruited in 2011. Every month for eight months, respondents were invited per email to answer questionnaires on different topics. The panel, GESIS Online Panel Pilot, is described in more detail in Chapter 2. Chapter 1 provides an overarching theoretical framework and the outline of this dissertation.

The goal of Chapter 2 was to answer the question: How good are the final estimates collected in the online panel? The evaluation is performed by comparing and contrasting the estimates from the online panel to the subsamples of Internet users from two high-quality face-to-face reference surveys: the German General Social Survey "ALLBUS" and the German sample of the European Social Survey (ESS). Furthermore, since researchers may not only be interested in single estimates, we study to what extent the survey results are comparable when the data are used for modeling social phenomena. We found differences among the surveys on most of the socio-demographic and attitudinal variables. These differences, however, average to a few percentage points. The regressions coefficients showed the same direction in most cases but differed in magnitude. In Chapter 2 we also perform post-survey adjustments. Nonprobability online panels often claim to achieve representativeness for the target population by means of weighting, that is, "rebalancing" their samples to match the (demographic) characteristics of the general population. Although our panel was based on a probability sample, persons who did not use the Internet were excluded by design. However, post-stratification weighting did not bring the estimates from the online panel closer to the estimates from the two reference surveys.

The next three chapters concentrated on the processes that can cause errors in the final data collected in the online panel. The focus of Chapter 3 was nonresponse. The recruitment for our panel included a short telephone interview at the end of which respondents were asked if they were willing to participate in an online panel. Persons, who agreed and provided their email addresses in order to be sent an email invitation to the survey, could fail to start the online survey. These two additional stages of expressing willingness to participate and actually starting the survey might aggravate the initial nonresponse error in the recruitment interview if respondents systematically differ from nonrespondents. Using the information collected during the telephone interview, we found that both groups (persons willing to participate and persons starting the online survey) are selective. They differed in demographic and attitudinal characteristics, but more importantly in their experience with the Internet and with online surveys. Those persons who used the Internet less frequently as well as those who had not completed any online surveys in the past were less likely to be willing to participate and less likely to start the online survey even if they agreed to join the panel. Moreover, we found that incentives (rewards for participation) played an important role as did the fieldwork agencies that performed the recruitment.

Chapter 4 also concentrated on nonresponse, but the focus was on nonresponse which occurred after study participants have completed at least one online questionnaire (attrition). Nonresponse which accumulates in the course of the panel can negatively affect representativeness of the online panel. Furthermore, it diminishes the sample size, which makes the estimates less precise. Therefore, it is important to secure continuous participation of respondents. There are many reasons why people choose to continue participation in the panel (panel wave). For example, respondents may enjoy completing the surveys, feel that by answering surveys they contribute to society, or participate out of habit etc. The incentives (rewards for participation, often monetary) may also play an important role. In this chapter we contrasted the role that incentives and non-reward motivation play for respondent attrition. We looked at how respondents' evaluation of the questionnaire measured in each wave was related to attrition in contrast to incentives. We found that respondents who viewed the survey as burdensome (long, difficult, too personal) were more likely to attrite. Incentives did not compensate for this burdensome experience, although in general receiving incentives and the incentive amount were negatively related to attrition. In this chapter we also compared the participation patterns and found that respondents who attritted monotonically and respondents who participated only in the first wave are more likely to attrite after a burdensome experience. However, the group of respondents which showed an irregular participation pattern was comparable in its characteristics to respondents who completed every wave.

Another possible drawback of panel surveys, in addition to panel attrition, is panel conditioning. Panel conditioning refers to the learning effect that occurs as a result of participating in previous panel waves. Panel conditioning may occur as a result of learning the questionnaire content or learning the interviewing rules by the respondents. In this chapter the latter is studied. Learning the rules of the interview may be advantageous for data quality if panel members provide more accurate responses to questions because of increased familiarity with a survey procedure and their tasks as respondents. However, it may also be disadvantageous, for example, if respondents learn to recognize filters (questions that are followed by a series of follow-ups) or repeated elements within the questionnaire and take "shortcuts", that is, to answer in a manner that avoids the follow-up questions. In this case, the data provided by respondents will be biased. In Chapter 5 both advantageous and disadvantageous conditioning are studied. We conducted two experiments that involve rotating the questionnaires to find out if the group of respondents who are longer with the panel would answer differently than the group of respondents with a shorter duration. We found limited evidence of advantageous conditioning and no evidence of disadvantageous conditioning. On the contrary, both groups of respondents were motivated respondents. One of the problems in studying panel conditioning is that conditioning effects need to be separated from attrition. In this chapter, we use propensity-score weighting to adjust for attrition and other confounding factors. We rely on the analysis of paradata, that is, data about the question-answering process gathered automatically by the software, to rule out alternative explanations of panel conditioning.

In the final chapter of this dissertation, we again concentrated on the estimates. The data from various waves of the online panel was compared to the two reference surveys: the German General Social Survey "ALLBUS" from the year 2010 and from the year 2012. Both face-to-face surveys were recruited by the same fieldwork agency and employed an almost identical design. Therefore, they together can serve as a "reference mode" if we want to compare them with the online panel. We studied mode system effects, that is, the differences in the estimates as the results of the whole process by which they were collected. We used questions with identical wordings that were present in both ALLBUS surveys and replicated in the online panel. These questions were asked from the respondents of the online panel only once, therefore, they could not be affected by panel conditioning. The differences in sample compositions among the surveys were adjusted by propensity-score weighting. We found that all three surveys differed from each other on attitudinal measures. However, for factual questions the two face-to-face surveys differed from the online panel and not from each other. We interpreted this finding as evidence of a mode system effect. If we had three surveys that varied in fieldwork procedures, differences could have been found among all three of them. In this study, the stable measures (factual questions) showed that one system (ALLBUS surveys) produced different results than the other system (online panel). Had we only taken the attitudinal questions, we would not have observed this effect. This chapter showed the importance of having two reference surveys that employed almost identical recruitment protocols and interviewing procedures. With respect to the differences in estimates, we found these differences to be small in magnitude. Our overall conclusion is that the data from the online panel is fairly comparable to the data from high-quality face-to-face surveys.

There are topics in online panel data quality which were not covered in this dissertation. Among others, these include the operational issues of panel recruitment and maintenance, the design of the online questionnaires, completion of online questionnaires on mobile devices and related measurement issues, and the interplay between survey errors. Nonetheless, the results of this dissertation provide additional insight into the processes that contribute to the quality of the final data produced in a probability-based panel with a typical recruitment process. These results can guide researchers who plan to build probability-based online panels of Internet users or of the general population. The results of the specific chapters (e.g., Chapter 2 and Chapter 6) might prove useful for existing panels that consider switching to the online mode.

Nederlandse samenvatting (Summary in Dutch)

Een survey is een onderzoeksmethode methode om op een systematische wijze informatie te verzamelen van een steekproef van entiteiten, gewoonlijk personen, om deze te generaliseren naar de gehele populatie. Informatie verzameld in survey–onderzoek wordt voor verschillende doelen gebruikt. Onderzoekers zoals sociologen, politicologen en economen gebruiken deze informatie om sociale fenomenen en maatschappelijke processen te onderzoeken. Politici gebruiken de resultaten van surveys om beleidsbeslissingen te ondersteunen. Commerciële bedrijven gebruiken de informatie verzameld in surveys om de tevredenheid of voorkeuren van de klanten te meten met het doel om producten en diensten verder te ontwikkelen. Verkiezingsonderzoek (poll) is een speciale vorm van surveyonderzoek met het doel het publiek en de politieke kandidaten te informeren over stemintenties, verwachte verkiezingsuitslagen, en belangrijke politiek punten.

De informatie verzameld in surveys moeten van hoge kwaliteit zijn om ervoor te zorgen dat organisaties en individuen die deze informatie gebruiken (datagebruikers) juiste beslissingen maken. Dit betekent allereerst dat de organisaties die surveys uitvoeren (dataproducenten) ervoor moeten zorgen dat de data geen fouten bevat. Aan de andere kant hebben dataproducenten een beperkt budget en beperkte middelen, dus ze moeten de kwaliteit maximaliseren met deze beperkingen.

Surveys die worden uitgevoerd via het internet kosten minder geld vergeleken met surveys die meer traditionele manieren (modes) van datacollectie hanteren (face-to-face interviews, telefonische interviews, en postale enquêtes). Face-to-face en telefoon surveys worden uitgevoerd door interviewers, dus internet surveys hebben het voordeel dat er geen interviewers betaald hoeven te worden. Surveys via de post hebben te maken me print- en portokosten en met de kosten van het verwerken en invoeren van de teruggestuurde enquêtes, terwijl in online surveys de data die door de respondenten wordt verstrekt direct in elektronische vorm beschikbaar is. Omdat het percentage van inwoners met internettoegang in veel landen sterk toeneemt zijn online surveys, en specifiek online panel surveys (surveys waarin respondenten frequent worden gevraagd een survey in te vullen in gespecificeerde tijdsintervallen), een lucratief alternatief voor surveys uitgevoerd in meer traditionele modes.

Sommige kenmerken van een online panel survey (online panels) kunnen echter verschillende problemen veroorzaken, die de datakwaliteit kunnen aantasten. De belangrijkste vraag die deze dissertatie tracht te beantwoorden is hoe goed de kwaliteit van de data verzameld in de online panels is. Er zijn twee belangrijke typen online panels: panels waarin iedereen mee mag doen (vrijwillige, online panels die niet gebaseerd zijn op een waarschijnlijkheidssteekproef) en panels waarbij de respondenten worden geselecteerd door de onderzoekers als een gevolg van een statistische procedure van een aselecte steekproef (online panels, gebaseerd op een waarschijnlijkheidssteekproef). Bij het laatste type kunnen valide conclusies worden getrokken over de populatie waarin men is geïnteresseerd. Deze dissertatie concentreert zich op aselect, via een waarschijnlijkheid steekproef samengestelde, online panels.

&

Er moeten verschillende stappen worden genomen om individuen te werven voor een aselect online panel. Allereerst moet de onderzoeker de populatie definiëren die hij of zij wil onderzoeken en een steekproefkader vinden. Voor een face-to-face survey of een survey via de post kan een populatieregister als steekproefkader worden gebruikt waaruit deelnemers aan de studie aselect kunnen worden getrokken. Echter, zulke registers (als ze beschikbaar zijn) bevatten geen e-mailadressen. Daarom moeten geselecteerde deelnemers worden gecontacteerd in een alternatieve mode (face-to-face of per post). Een vergelijkbare procedure is ook telefonisch mogelijk, via Random Digit Dialing, en in sommige landen is een steekproef gebaseerd op adressen is een alternatief voor populatie registers. Niet iedereen in de populatie heeft toegang tot of gebruikt het Internet. Wanneer een onderzoeker de hele populatie wil onderzoeken, en niet alleen de internetgebruikers, zou het uitsluiten van de populatie die geen toegang heeft tot het Internet tot dekkingsfouten kunnen leiden. Dekkingsfouten zullen voorkomen wanneer internetgebruikers verschillen van niet-internetgebruikers in aspecten die belangrijk zijn voor de onderzoeker. Om dit te voorkomen zouden niet-Internet gebruikers een alternatieve dataverzamelingsmethode aangeboden kunnen krijgen om de survey te voltooien, een andere mogelijkheid is om de benodigde apparatuur (computer, internet) die nodig is om te participeren aan te bieden.

Ook in de volgende stap kunnen problemen ontstaan wanneer personen die zijn geselecteerd voor de studie niet kunnen worden bereikt (noncontact), of als zij wel bereikt kunnen worden, zij niet kunnen of willen deelnemen aan de studie (nonrespons). Wanneer deze personen systematisch verschillen van diegenen die wel deelnemen aan de studie, zal dit resulteren in nonresponsfouten. In panel surveys is de situatie nog gecompliceerder omdat deelnemers aan de studie die in eerste instantie toe hebben gezegd deel te nemen ervoor zouden kunnen kiezen dit op latere meetmomenten niet meer te doen (paneluitval). Daarnaast zouden respondenten door frequente deelname kunnen leren hoe zij (de lengte van) de vragenlijst kunnen manipuleren (bijvoorbeeld door onwaar, negatief op filtervragen te antwoorden) om de belasting van deelname te verlichten. Dit zal een negatieve invloed hebben op de metingen (meetfouten). Alle bovengenoemde fouten kunnen opeenstapelen, en de schattingen (het eindproduct van een survey) en de conclusies die op basis hiervan worden getrokken, kunnen daardoor misleidend zijn. De hierboven beschreven fouten zijn niet uniek voor online panels. Zij komen ook voor bij cross-sectionele (eenmalige) surveys of panel surveys uitgevoerd in andere modes. Er zijn echter wel een aantal specifieke problemen bij online onderzoek, bijvoorbeeld, wanneer een geselecteerde respondent weigert deel te nemen vanwege onvoldoende vaardigheden of ervaring met het Internet.

Die vijf studies in deze dissertatie richten zich op vragen rondom datakwaliteit. Elk hoofdstuk is een aparte studie. Sommige hoofdstukken concentreren zich op de kwaliteit van schattingen, terwijl andere hoofdstukken de nadruk ligt op de processen die tot fouten in schattingen kunnen leiden. Data van een aselect online panel van internetgebruikers die via de telefoon zijn geworven in Duitsland (GESIS Online Panel Pilot) wordt gebruikt in alle studies. De respondenten in de GESIS Online Panel Pilot zijn geworven in 2011. De respondenten werden gedurende acht maanden lang iedere maand met een e-mail uitgenodigd om vragenlijsten over verschillende onderwerpen te beantwoorden. De -in het GESIS Online Panel Pilot project- verzamelde data wordt uitgebreid beschreven in hoofdstuk 2.

Hoofdstuk 1 presenteert een overkoepelend theoretisch kader en een overzicht van de hoofdstukken in deze dissertatie. Het doel van hoofdstuk 2 was om de vraag te beantwoorden: hoe goed zijn de uiteindelijke schattingen verkregen via het online panel? De evaluatie is uitgevoerd door het vergelijken en contrasteren van schattingen van het online panel met deelsteekproeven van internetgebruikers uit twee face-to-face referentiesurveys: de German General Social Survey “ALLBUS” en de Duitse steekproef van de European Social Survey (ESS). Omdat onderzoekers niet alleen geïnteresseerd zijn in individuele schattingen, onderzoeken we in welke mate de resultaten vergelijkbaar zijn wanneer de data wordt gebruikt om sociale fenomenen te modelleren. We hebben verschillen gevonden tussen de surveys voor de meeste van de sociaal-demografische en attitude variabelen. Deze verschillen zijn doorgaans echter maar enkele procentpunten. De regressiecoëfficiënten lieten in de meeste gevallen dezelfde richting zien maar verschilden in grootte. In hoofdstuk 2 voeren we ook post-survey correcties uit. Selecte online panels (niet gebaseerd op een waarschijnlijkheidssteekproef) claimen vaak representatief te zijn voor de doelpopulatie door weging, dat is het “herbalanceren” van de steekproeven om overeen te komen met de (demografische) karakteristieken van de algemene populatie. Ondanks dat ons panel op een aselecte steekproef was gebaseerd, zijn personen die geen internet gebruiken uitgesloten van deelname door het design. Weging door poststratificatie brengt echter de schattingen van het online panel niet dichter bij de schattingen van de twee referentiesurveys.

De volgende drie hoofdstukken zijn toegespitst op de processen die fouten kunnen veroorzaken in de uiteindelijke data verzameld in het online panel. De focus van hoofdstuk 3 was nonrespons. Bij het werven van ons panel werd er een kort telefonisch interview gehouden waarbij aan het einde werd gevraagd of respondenten wilden deelnemen aan een online panel. De respondenten die hiermee instemden en hun e-mailadres gaven, zodat zij via e-mail een uitnodiging voor de survey konden ontvangen, hebben desondanks niet allen aan de eerste online survey deelgenomen. Deze twee extra fases, het toezeggen om mee te werken aan de survey en het daadwerkelijk starten van de online vragenlijst, zouden de initiële nonresponsfout in het wervingsinterview kunnen vergroten, wanneer respondenten systematisch verschillen van nonrespondenten. Gebruikmakend van de in het telefonische interview verzamelde informatie, vonden we dat beide groepen (personen bereid om te participeren en personen die de online vragenlijst daadwerkelijk starten) selectief zijn. Ze verschilden in demografische en gedragskenmerken, maar nog belangrijker, in hun ervaring met het Internet en met online surveys. Zowel degenen die het internet minder vaak gebruikten als degenen die in het verleden geen online vragenlijst hadden beantwoord, waren minder geneigd om deel te nemen en de online survey te starten. Zelfs wanneer ze hadden toegezegd om aan het panel deel te nemen. Bovendien vonden we dat beloningen voor deelname een belangrijke rol speelden. Dit gold ook voorde veldwerkbureaus die de werving uitvoerden

Hoofdstuk 4 concentreert zich ook op nonrespons, maar de focus ligt nu op nonrespons die voorkwam nadat deelnemers tenminste één online vragenlijst hadden voltooid (paneluitval).

&

Nonresponse die zich opstapelt in de loop van het panel kan een negatieve invloed hebben op de representativiteit van het online panel. Bovendien verkleint het de steekproefgrootte, wat ertoe leidt dat de schattingen minder precies zijn. Daarom is het belangrijk ervoor te zorgen dat er een continue deelname van respondenten is. Er zijn vele redenen waarom mensen besluiten deel te blijven nemen aan het panel. Respondenten kunnen het bijvoorbeeld leuk vinden een vragenlijst in te vullen, het gevoel hebben dat zij door het beantwoorden van een survey bijdragen aan de maatschappij, of uit gewoonte deelnemen etc. De beloningen voor deelname, vaak in de vorm van geld, kunnen ook een belangrijke rol spelen. In dit hoofdstuk contrasteren we de rol die beloningen en intrinsieke motivatie spelen in nonrespons. We vergeleken hoe de evaluatie van de vragenlijst door respondenten (die in iedere panelwave werd gemeten) en beloningen zijn gerelateerd aan panel attritie. Hier kwam naar voren dat respondenten die de survey belastend vonden (lang, moeilijk, te persoonlijk) meer geneigd waren tot uitvallen. Beloningen compenseerden dit niet, hoewel in het algemeen het ontvangen van beloningen en de grootte van de beloning negatief waren gerelateerd aan paneluitval. In dit hoofdstuk hebben we ook participatiepatronen vergeleken en gevonden dat respondenten die monotoon uitvielen en respondenten die alleen in de eerste wave deelnamen meer geneigd zijn uit te vallen na een belastende ervaring.

Een ander mogelijk nadeel van panel surveys, naast uitval, is panel conditionering. Panel conditionering verwijst naar het leereffect dat optreedt als gevolg van het deelnemen aan eerdere panelwaves. Panel conditionering kan optreden als gevolg van het leren van de vragenlijstinhoud of het leren van de interviewregels door de respondenten. In dit hoofdstuk wordt het laatste onderzocht. Het leren van de regels van het interview zou ten goede kunnen komen aan de datakwaliteit als panelleden nauwkeuriger antwoorden geven op vragen door de toegenomen vertrouwdheid met een surveyprocedure en hun taken als respondent. Het zou echter ook nadelig kunnen zijn, bijvoorbeeld wanneer respondenten filters (vragen die worden gevolgd door een serie van follow-upvragen) of herhaalde elementen in de vragenlijst leren herkennen en "kortere wegen" nemen, dat wil zeggen, antwoorden geven op zo'n manier dat geen follow-upvragen krijgen. In dit geval zal de data verstrekt door de respondenten vertekend zijn. In hoofdstuk 5 worden zowel voordelige als nadelige conditioneringen bestudeerd. We hebben twee experimenten uitgevoerd waarbij vragenlijsten werden omgedraaid om te onderzoeken of de groep respondenten die al langer in het panel zat anders zou antwoorden dan de groep respondenten die korter in het panel zat. Er werd beperkt bewijs gevonden voor voordelige conditioneringen en geen bewijs voor nadelige conditioneringen. Echter, beide groepen bestonden uit gemotiveerde respondenten. Een van de problemen bij het onderzoeken van panel conditionering, is dat conditioneringseffecten onderscheiden moeten worden van uitvaleffecten. In dit hoofdstuk gebruiken we propensity-scores om te wegen en op die manier te corrigeren voor uitval en andere confounding factoren. We gebruiken paradata, data over het proces van het beantwoorden van de vragen die automatisch wordt verzameld door de software, om alternatieve verklaringen van panel conditionering uit te sluiten.

In het laatste hoofdstuk van deze dissertatie concentreren we ons weer op de schattingen. De data van verscheidene waves van het online panel wordt vergeleken met data uit twee referentiesurveys: de German General Social Survey "ALLBUS" 2010 en 2012. Beide face-to-face surveys zijn door hetzelfde veldwerkbureau geworven en gebruikten een vrijwel identiek design. Daarom vormen zij samen de "referentiemode" voor een vergelijking met het online panel. We hebben mode systeemeffecten bestudeerd, dat wil zeggen, de verschillen in de schattingen als een resultaat van het hele proces waarin zij werden verzameld. We gebruikten vragen met identieke bewoordingen die in beide ALLBUS surveys aanwezig waren en hebben deze herhaald in het online panel. Deze vragen werden slechts eenmaal aan de respondenten in het online panel gesteld en daarom konden zij niet worden beïnvloed door panel conditionering. De verschillen in steekproefsamenstellingen tussen de surveys werd gecorrigeerd door propensity-score weging. We vonden dat alle drie de surveys van elkaar verschilden op attitudemetingen. Voor feitelijke vragen verschillen de twee face-to-face referentie surveys echter van het online panel en niet van elkaar. We hebben deze bevinding geïnterpreteerd als bewijs voor een mode-systeemeffect. Als we drie surveys uitvoeren die verschillen in veldwerkprocedures, kunnen er in principe verschillen worden gevonden tussen al deze drie surveys. In deze studie lieten de stabiele metingen (feitelijke vragen) zien dat één systeem (de twee ALLBUS surveys) resultaten produceerden die verschillen van het andere systeem (het online panel). Wanneer we alleen naar attitudevragen hadden gekeken hadden we dit effect niet geobserveerd. Dit hoofdstuk laat het belang zien van het hebben van twee referentiesurveys die vrijwel identieke wervings- en interviewprocedures hanteren. Met betrekking tot de verschillen in schattingen vonden we dat deze klein waren. Onze algemene conclusie luidt dat de data van het online panel tamelijk vergelijkbaar is met de data van face-to-face surveys van hoge kwaliteit.

Er zijn onderwerpen in online panel datakwaliteit die niet worden behandeld in deze dissertatie. Deze omvatten onder andere de operationele problemen van panelwerving en panelonderhoud, het design van online vragenlijsten, het voltooien van online vragenlijsten op mobiele apparaten en hieraan gerelateerde meetproblemen, en de wisselwerking tussen surveyfouten. Desalniettemin, geven de resultaten van deze dissertatie inzicht in de processen die bijdragen aan de kwaliteit van de uiteindelijke data die wordt geproduceerd in een op een kanssteekproef gebaseerd panel van internetgebruikers of van de algemene populatie. Deze resultaten zijn bruikbaar voor onderzoekers die een online panel willen opzetten gebaseerd op een kanssteekproef. De resultaten van specifieke hoofdstukken (bijvoorbeeld hoofdstuk 2 en hoofdstuk 6) kunnen nuttig zijn voor bestaande panels die overwegen om om te schakelen naar een online mode.

About the author

Bella Struminskaya (1985) studied sociology and economics at Novosibirsk State University (Russia) and Philipps University of Marburg (Germany). She holds a BA in Sociology from Novosibirsk State University and an MA in Sociology from the University of Mannheim (Germany). In August 2010, Bella started working as a researcher at the department of Survey Design and Methodology at GESIS – Leibniz Institute for the Social Sciences in Mannheim.

The five studies in this dissertation, which was supervised by Prof. Dr. Edith de Leeuw (Utrecht University) and Dr. Lars Kaczmirek (GESIS), are the result of Bella's work in the project GESIS Online Panel Pilot led by Dr. Lars Kaczmirek. Apart from the studies in this dissertation, Bella has published on optimization of RDD mobile phone surveys in *Survey Practice* as a result of this project.[39]

In 2013, Bella started working in the project GESIS Panel (a probability-based mixed-mode access panel for the social sciences in Germany), for which the GESIS Online Panel Pilot was a predecessor. Her research interests include Internet surveys and online panels, mixed-mode designs, and survey operations.

&

[39] Struminskaya, B., Kaczmirek, L., Schaurer, I., Bandilla, W., Gabler, S. & Häder, S. (2011). Identifying Non-working Numbers in Cell Phone RDD Samples via HLR-Lookup Technology: Reduced Survey Costs and Higher Precision in Response Rate Calculation. *Survey Practice,* (August). http://surveypractice.wordpress.com/2011/08/30/identifying-nonworking-numbers/