mda
methods, data, analyses
Journal for Quantitative Methods and Survey Methodology
Volume 11, 2017 | 2

Edited by Annelies G. Blom, Edith de Leeuw, Gabriele Durrant, Bärbel Knäuper

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

Editors: Annelies G. Blom (Mannheim, editor-in-chief), Edith de Leeuw (Utrecht), Gabriele Durrant (Southampton), Bärbel Knäuper (Montreal)

Advisory board: Hans-Jürgen Andreß (Cologne), Andreas Diekmann (Zurich), Udo Kelle (Hamburg), Dagmar Krebs (Giessen), Frauke Kreuter (Mannheim), Norbert Schwarz (Los Angeles), Christof Wolf (Mannheim)

Managing editor: Sabine Häder
GESIS – Leibniz Institute for the Social Sciences
PO Box 12 21 55
68072 Mannheim
Germany
Tel.: + 49.621.1246282
E-mail: mda@gesis.org
Internet: www.mda.gesis.org

Methods, data, analyses (mda) publishes research on all questions important to quantitative methods, with a special emphasis on survey methodology. In spite of this focus we welcome contributions on other methodological aspects. We especially invite authors to submit articles extending the profession's knowledge on the science of surveys, be it on data collection, measurement, or data analysis and statistics. We also welcome applied papers that deal with the use of quantitative methods in practice, with teaching quantitative methods, or that present the use of a particular state-of-the-art method using an example for illustration.

All papers submitted to mda will first be screened by the editors for general suitability and then double-blindly reviewed by at least two reviewers. Mda appears in two regular issues per year (January and July).

Print: Bonifatius Druck GmbH Paderborn, Germany

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

© of the compilation GESIS, Mannheim, July 2017

All content is distributed under the terms of the Creative Commons Attribution 3.0 License. Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.

Content

Editorial

This issue of mda takes a special place in the hearts of the mda editorial team. The reason is that, in addition to three fine research reports by Melanie Revilla (RECSM – Universitat Pompeu Fabra, Spain), Uta Landrock (University of Kaiserslautern, Germany), and Taylor Lewis (U.S. Office of Personnel Management, United States), we were delighted to receive and ultimately accept a review article by two most distinguished survey methodologists: Eleanor Singer and Mick P. Couper (Survey Research Center, Institute for Social Research, University of Michigan, United States).

It is with even greater sadness then that we learned that Eleanor Singer died aged 87 on 3rd June 2017. Eleanor was a highly respected researcher, who inspired generations of survey methodologists in the U.S., Europe, and worldwide.

The connecting theme of her research was the validity of measurements – that is, whether issues such as privacy and confidentiality, informed consent procedures, or incentives have an effect on survey participation, bias, and the accuracy of survey response data. Her studies adapted to technological advances in survey settings over the years, and explored the implications of face-to-face, mail, phone, and online surveys.

ISR Director David Lam: "Eleanor was a major figure in the field of survey methodology and she will be greatly missed by all who knew her. We are fortunate that she spent the last decades of her illustrious career at ISR, where she made major contributions to research, training, and the intellectual life of the Institute."

Bob Groves: "Eleanor was one of those productive scientists who was also an incredible magnet for collaboration. She ended up collaborating with half of the people in the building, was known as a wonderful mentor, and an exquisite writer. Whenever I would get back articles I submitted to her that she had rewritten, I realized she made my pieces better. As a collaborator you would discover that again and again."

Michael Traugott: "Eleanor was editor of Public Opinion Quarterly at a time when survey research and public opinion research became established in the university setting. By her selection of content and manuscripts, she – in a very important but subtle way – promoted and encouraged the study of academic survey methods as well as the current state of knowledge that is very important to survey researchers."

Eleanor Singer was born in Vienna, Austria, in 1930. When she was 8 years old, her family fled the rise of Nazi Germany in Europe and settled in Astoria, New York. She completed a B.A. in English at Queens College in 1951, where she met her late husband Alan Singer. In her early career, Singer worked as a book editor

at various publishing houses, including Teachers College Press, and increasingly specialized in books about social science.

During this time, survey research, public opinion, and polling began to grow as a field of study in the United States. Singer developed an interest in sociology – in particular surveys and survey research – and decided to pursue graduate school at Columbia University in 1959. She earned a Ph.D. in Sociology in 1966.

During the course of her studies, she met and worked with illustrious mentors including Paul Lazarsfeld and Robert Merton, and her dissertation sponsor Herbert H. Hyman, who introduced Singer to public opinion research and survey methodology. She went on to conduct research at Columbia University and University of Chicago, and worked as a social science analyst at the U.S. Bureau of the Census.

In her distinguished career she received numerous appointments and distinctions, including the Monroe G. Sirken Award in Interdisciplinary Survey Methods Research for "significant contributions in our understanding of survey participation, sources of nonresponse bias, and factors affecting survey responses; for pioneering research on the use and effects of incentives; and for leadership in developing awareness and understanding of ethical issues in survey research."

I had the honor to meet Eleanor and discuss research with her during a research visit at SRC in June 2013. I was impressed by her inquisitive mind, passion for knowledge and research, humbleness and kindness. She will be sorely missed by colleagues, friends and family.

I borrowed liberally from a more detailed obituary on the ISR website:
http://home.isr.umich.edu/releases/survey-researcher-eleanor-singer-dies

Annelies G. Blom
Editor-in-Chief

methods, data, analyses | Vol. 11(2), 2017, pp. 115-134 DOI: 10.12758/mda.2017.01

Some Methodological Uses of Responses to Open Questions and Other Verbatim Comments in Quantitative Surveys

Eleanor Singer & Mick P. Couper
Survey Research Center, University of Michigan

Abstract

The use of open-ended questions in survey research has a very long history. In this paper, building on the work of Paul F. Lazarsfeld and Howard Schuman, we review the methodological uses of open-ended questions and verbatim responses in surveys. We draw on prior research, our own and that of others, to argue for increasing the use of open-ended questions in quantitative surveys. The addition of open-ended questions – and the capture and analysis of respondents' verbatim responses to other types of questions – may yield important insights, not only into respondents' substantive answers, but also into how they understand the questions we ask and arrive at an answer. Adding a limited number of such questions to computerized surveys, whether self- or interviewer-administered, is neither expensive nor time-consuming, and in our experience respondents are quite willing and able to answer such questions.

Keywords: open questions; textual analysis; verbatim comments

© The Author(s) 2017. This is an Open Access article distributed under the terms of the Creative Commons Attribution 3.0 License. Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.

1 Introduction

More than 75 years ago Lazarsfeld (1935), in "The Art of Asking Why," offered advice on the proper (and improper) deployment of open-ended questions. He identified six main functions of the open-ended interview: clarifying the meaning of a respondent's answer, singling out the decisive aspects of an opinion, discovering what has influenced an opinion, determining complex attitude questions, interpreting motivations, and clarifying statistical relationships. In "The Controversy over the Detailed Interview – An Offer for Negotiation," prepared in response to an invitation to adjudicate professional disagreements over the relative merits of closed versus open-ended questions, he argued that both open and closed questions should be used in a comprehensive research program (Lazarsfeld, 1944).

Over time, the economics of survey research gradually drove out open-ended interviewing as a technique for quantitative large-scale studies (cf. Geer, 1991). But about a quarter century later Howard Schuman proposed an ingenious solution to the cost dilemma. In "The Random Probe" (1966), he pointed out that most of the functions of open-ended questions noted by Lazarsfeld could, in fact, be fulfilled by probing a randomly selected subset of responses to closed-ended questions with open-ended follow-ups. Such probes could be used to clarify reasons for the response, clear up ambiguities, and explore responses that fell outside the expected range of answers. Because they would be put only to a subset of respondents, they would reduce the cost of recording and coding; but since the subsample was randomly selected, the results could be generalized to the sample as a whole. Schuman himself has made much use of this technique over his long career in survey research, reprised in his most recent book, *Meaning and Method* (2008). Nevertheless, the promise of this approach has not yet been fully realized, despite the development of technologies that make it even easier to implement today.

Here, we review several primarily methodological uses of open-ended questions and give examples drawn from our own research as well as that of others. We believe the adaptation of open-ended questions to some functions in quantitative surveys for which they have not previously been used, or used only rarely, will result in more respondent-focused surveys and more accurate and useful data. The paper argues for more inclusion of open-ended questions in quantitative surveys and discusses the technological and methodological advances that facilitate such inclusion. The major advantage of embedding such questions in actual surveys rather than restricting their use to qualitative interviews is the breadth and representativeness of coverage they provide at little additional cost. Such use should

Direct correspondence to
Mick P. Couper, ISR, University of Michigan, P.O. Box 1248, Ann Arbor, MI 48106, U.S.A.
E-mail: mcouper@umich.edu

complement, not replace, the use of open questions and verbatim responses during the instrument development and pretesting process.

We take a broad perspective on open questions in this paper, including any question where the respondent's answers are not limited to a set of predefined response options. Couper, Kennedy, Conrad, and Tourangeau (2011) review different types of such responses, including questions eliciting narrative responses (e.g., "What is the biggest problem facing the country today?") and those soliciting a numeric response (e.g., "During the past 12 months, how many times have you seen or talked with a doctor about your health?"). We include all these types, and expand the notion to include verbatim responses to closed questions that do not fall within the prescribed set of response alternatives.

2 Why Add Open-Ended Questions to Surveys?

As already noted, Schuman (1966) proposed following some closed questions with open-ended probes administered to a random sample of respondents in order to clarify their answers and – which is often forgotten – to establish the validity of closed questions (Schuman & Presser, 1979). We believe such probes can serve a number of other important functions as well. For all of these, embedding the probes in ongoing surveys has clear benefits. First, there is a good chance of capturing the full range of possible responses, since the survey is administered to a random sample of the target population; and second, if the survey is web-based or administered by an interviewer using a computer, the responses can be captured digitally, facilitating automatic transcription or computer-assisted coding, in turn reducing the cost and effort involved in analyzing the responses. Such "random probes" thus provide a useful addition, and in some cases an alternative, to a small number of qualitative interviews administered to convenience samples.

In what follows, we identify seven primarily methodological uses of open-ended questions: Understanding reasons for reluctance or refusal; determining the range of options to be used in closed-ended questions; evaluating how well questions work; testing methodological theories and hypotheses; checking for errors; encouraging more truthful answers; and providing an opportunity for feedback. We omit another frequent use of open-ended questions – namely, as an indicator of response quality (e.g. Galesic & Bosnjak, 2009; for a summary of this use of open-ended questions in incentive experiments see Singer & Kulka, 2002).

2.1 Understanding Reasons for Refusal

The first use of open responses lies outside the traditional domain of standardized survey instruments. Introductory interactions were long thought of as something external to the survey itself, and therefore as something not subject to systematic measurement. However, the early pioneering work of Morton-Williams (1993; see also Morton-Williams & Young, 1987) showed that systematic information can be collected about these interactions and used for quantitative analysis, and a few studies have collected systematic data about "doorstep interactions" between interviewers and respondents in an effort to use respondent comments to predict the likelihood of response and allow interviewers to "tailor" their comments to specific respondent concerns (Morton-Williams & Young, 1987; Morton-Williams, 1993; Groves & Couper, 1996; Campanelli et al., 1997; Couper, 1997; Sturgis & Campanelli, 1998; Groves & McGonagle, 2001; Couper & Groves, 2002; Bates et al., 2008).

In an early paper, Couper (1997) demonstrated that there is some veracity to the reasons sample persons give for not wanting to participate in a survey. Those who say "not interested" did indeed appear to be less interested, engaged, and knowledgeable about the topic (elections) than those (for example) who gave "too busy" as a reason. Interviewer observations are now a standard part of many survey data collection protocols. Often the verbatim reactions of householders to the survey request are field-coded by interviewers. Recent efforts have focused on improving the quality of such observations (see, e.g., West, 2013; West & Kreuter, 2013, 2015).

For example, the US Census Bureau makes data from its contact history instrument (CHI; see, e.g., Tan, 2011), which systematically captures information on interviewer-householder interactions, available to researchers. The CHI provides information about the characteristics of all sample members with whom contact was made, permitting not only the tailoring of subsequent contacts to counteract reservations that may have been expressed at the prior encounter, but also to predict what kinds of responses are likely to lead to final refusals and which are susceptible of conversion. Bates, Dahlhamer, and Singer (2008), for example, analyzed the effect of various respondent concerns, expressed during a personal contact with an interviewer, on cooperation with the National Health Interview Survey. While acknowledging various limitations of the CHI instrument, including the fact that recording and coding the concerns involve subjective judgments by interviewers as well as possible recall error if such concerns are not recorded immediately, the authors report a number of useful findings in need of replication. Thus, for example, although 23.9% of households claimed they were "too busy" to do the interview during at least one contact, 72.8% of households expressing this concern never refused and only 10.3% were final refusals. Similarly, although 13.3% of households expressed privacy concerns, 62.9% of those expressing privacy concerns never

refused, and only 13.9% were final refusals. On the other hand, 34.1% of those (12.7% of households) saying "not interested" and "don't want to be bothered" never became respondents *(ibid*., Table 1). Because interactions between interviewers and respondents were not recorded verbatim in this study, we can only surmise why certain concerns were more amenable to mitigation than others, or guess at which interviewer conversational strategies might have been successful. While early methodological studies (most notably Morton-Williams, 1993) had interviewers tape-record the doorstep interactions, most subsequent work has required interviewers to report their observations of the interaction, a process subject to measurement error. Portable, unobtrusive digital recorders, increasingly an integral component of the laptop and tablet computers interviewers are using for data collection, make such doorstep recording increasingly feasible.[1] Recording of introductory interactions in telephone surveys is logistically even easier (e.g., Couper & Groves, 2002; Benki et al., 2011; Conrad et al., 2013).

Modes of interviewing that record the entire interaction, rather than manually recording only the respondent's concern, could begin to provide answers to questions relating to the process of gaining cooperation. For example, Maynard, Freese, and Schaeffer (2010) draw on conversation-analytic methods and research to analyze interviewer-respondent interactions in order to better understand the process of requesting and obtaining participation in a survey interview. The authors state, "This article contributes to understanding the social action of requesting and specifically how we might use insights from analyses of interaction to increase cooperation with requests to participate in surveys." Or, as the authors of the CHI paper note, "The potential of these new data to expand our understanding of survey participation seems great since they are collected at every contact, across modes, and across several different demographic surveys for which the US Census Bureau is the collecting agent." Indeed, they include an analysis of Consumer Expenditure Survey Data that replicates key findings of the main analysis (Bates et al., 2008).

2.2 Determining the Range of Options to Be Offered in Closed-Ended Questions

In "The Open and Closed Question", Schuman and Presser (1979) talk about the two main functions of open-ended questions: Making sure that all possible response options are included in the final questionnaire, and avoiding bias. They investigate experimentally how closely the coding of responses to an open-ended question replicates the *a priori* response alternatives assigned to a question about the importance of different aspects of work. Schuman has also talked about the

1 Note, however, that the technical developments do not address the informed consent issues raised by recording such introductory interactions.

importance of ascertaining the full range of response options to controversial questions before constructing a questionnaire. What, for example, is the most extreme response option to a question about the conditions under which abortion should be forbidden? Is it the termination of any pregnancy, however brief, or does it extend to the prevention of conception after unprotected intercourse, or even to the use of contraception? Schuman has suggested talking to groups holding extreme positions on both sides of a controversial issue before drafting questions about it. A possibly attractive alternative is to include the question in open-ended form – e.g., "What kinds of actions would you include in a definition of abortion?" – on a survey of a random sample of the target population which precedes the planned survey on abortion attitudes. Such a question should yield not only the extremes but also a distribution of intermediate responses. This is analogous to doing a small number of qualitative, semi-structured interviews prior to fielding a questionnaire, but has the advantage of doing so with a larger, more diverse sample in an ongoing survey at marginal cost. Behr et al. (2012, 2013, 2014) have investigated some factors contributing to the success of such probes in web surveys.

2.3 Evaluating How Well Questions Work

Just as open questions administered to a random sample can be useful in developing a questionnaire, so they can be useful in evaluating how well questions work in an actual survey. Martin (2004) discusses at length the use of open and closed debriefing questions administered after the main survey for evaluating respondents' understanding of key questions. Such questions have been used to measure the accuracy of respondents' interpretation of terminology, questions, or instructions; to gauge respondents' reactions or thoughts during questioning; and to obtain direct measures of missed or misreported information (e.g. Belson, 1981; DeMaio, 1983; DeMaio & Rothgeb, 1996; Oksenberg et al., 1991; Schuman, 1966). Hess and Singer (1995), for example, used open as well as closed questions administered to a random subsample of respondents to see how well respondents understood questions on a Food Insecurity supplement and how reliably some questions were answered.

Given the increasing ease with which digital recordings of the entire interview can be captured for analysis, verbatim responses to closed-ended questions in interviewer-administered surveys are becoming increasingly useful for evaluating the performance of survey questions. In the days of paper-and-pencil surveys, interviewers recorded the interviews on tape recorders. These were painstakingly coded and analyzed using methods such as behavior coding (see, e.g., Fowler & Cannell, 1996) or conversational-analytic methods (e.g., Schaeffer & Maynard, 1996; Maynard et al., 2002), often only in small pretests. Digital recordings integrated into computer-assisted interviewing (CAI) software make the task of finding responses to specific questions much easier. While much of the focus of this work has been on

evaluating interviewers, we believe such recordings are a valuable tool for evaluating survey questions. Indeed, Cannell and Oksenberg (1988) identified three main objectives of interview observation: 1) to monitor interviewer performance, 2) to identify survey questions that cause problems for the interviewer or respondent, and 3) to provide basic data for methodological studies.

To give one recent example: in the process of developing an online version of the Health and Retirement Study (see http://hrsonline.isr.umich.edu/) instrument, we were struggling with how to refer to family members (siblings or children) who had died since the last wave of data collection. HRS staff selected a number of recordings from the prior interviewer-administered wave of the survey where the data revealed a death of a sibling or child. By listening to these interactions, they were able to determine that the term "deceased" was used more frequently than "passed (away)" or other terms when referring to such family members. This enabled us to recommend appropriate wording for the online version of the survey.

Other examples of such targeted analysis include identifying questions with high rates of missing data to understand how respondents are communicating their responses; identifying concerns expressed about in-survey consent requests; understanding how respondents might qualify their answers in response to questions asking for exact qualities (e.g., income or assets, life expectancy probability, etc.); and the like. Both survey data and paradata can be used to identify questions for more detailed examination, whether qualitative or quantitative. We believe this is an under-utilized opportunity to use existing digital recordings to evaluate and improve survey questions.

2.4 Testing Methodological Theories and Hypotheses

Porst and von Briel (1995), Singer (2003), and Couper et al. (2008, 2010) have used open-ended questions in face-to-face, telephone, and online surveys to explore reasons people give for being willing (or unwilling) to participate in a hypothetical survey. Those who said they would be willing to participate cited things like wanting their opinions to be heard or wanting to contribute to the research goals, or their interest in the topic of the survey or the incentive associated with participation. Those who said they would not be willing to participate gave some general reasons – not interested, too long, too little time – as well as a large number of responses that were classified as privacy-related (e.g., Don't like intrusions; don't like to give financial information). A large number of responses pertained to survey characteristics, such as the topic or the sponsor, and a small number of comments indicated that respondents did not view the survey as offering enough benefits to make participation worthwhile.

These reasons can be reliably coded into a relatively small number of general categories – an egoistic-altruistic dimension (for example, "For the money," "To

help with the research"), another having to do with situational characteristics (for example, "I'm too busy," "I'm retired, so I have the time"), and still others having to do with characteristics of the survey ("It's too long," "I trust the sponsor"). Such categories could be used to develop a set of exhaustive, mutually exclusive reasons for (non)response, which in turn could be used to test hypotheses or theories about survey participation (Singer, 2011).

We have also asked respondents whether they would, or would not, be willing to permit researchers to make use of paradata – data automatically produced as a byproduct of answering survey questions on web-based surveys – both in connection with hypothetical vignettes and after completing an actual online survey (Couper & Singer, 2013; Singer & Couper, 2011), and followed this with open-ended questions about the reasons for their response. Exploratory questions about whether, and why, respondents would forbid or allow the use of paradata helped clarify the experimental results and can serve as the basis for subsequent quantitative surveys. For example, although we explained to respondents that we never track their browsing behavior, a large number of answers to open-ended questions referred to concerns about tampering with the respondent's computer, making clear that we had failed to reassure respondents on this point. Subsequent studies could test whether alternative reassuring messages are capable of reducing these concerns and increasing rates of participation. Recording and analyzing the responses given when respondents are asked for consent to linkage to administrative records (e.g., Sakshaug et al., 2012) or for physical or biomedical measurement (e.g., Sakshaug et al., 2010) could similarly help to identify and address reasons for non-compliance.

Examples also exist in other domains of the use of open-ended questions to aid in testing substantive or methodological hypotheses (our focus here being on the latter). For example, Yan, Curtin, and Jans (2010) used an open-ended question on income to measure trends in item nonresponse, which they hypothesized as being inversely related to trends in unit nonresponse. Mason, Carlson, and Tourangeau (1994) used an open-ended question to clarify the subtraction effect in answering part-whole questions. Tourangeau and colleagues (2014, 2016) used open-ended questions to understand the effect of using examples in survey questions.

2.5 Some Other Uses for Open-Ended Questions

In addition to those just discussed, we have found three other uses for open-ended questions. One relatively trivial use is as a check on the coding of the closed question that precedes the open-ended probe. In one particularly dramatic example drawn from our own research (Couper et al., 2008, 2010) we discovered, as a result of working with the open-ended responses, that the codes for answers to the question about willingness to participate had been reversed: Those who had said they would be willing to participate had been coded as if they would refuse, and vice

versa. Less dramatic examples occur more frequently: Someone who checks "9", indicating great willingness to participate, then enters a response to the open-ended probe that indicates the reverse – for example, "I probably wouldn't answer these kinds of questions in a face-to-face interview." It is then possible to correct the response to the closed question or, if the correct coding is not obvious, omit it altogether. Though most of the time they may not be worth the extra effort required, such checks can help to uncover problems with the closed question preceding the probe, and if even small errors cannot be tolerated, the effort may well prove worthwhile.

Still another function of open-ended questions appears to be to permit respondents to give more socially undesirable answers to threatening questions. This function was already pointed out by Blair et al. (1977) with respect to reports about sensitive behaviors such as the amount of alcohol drunk and the frequency of sexual intercourse. Compared with closed questions, open-ended questions elicited reports of a greater average number of drinks and more frequent sexual behaviors, whereas reports about non-sensitive behaviors, such as participation in sports, were not affected by the form of the question. In a subsequent study Tourangeau and Smith (1996) found that "responses to open-ended questions generally fell between responses to the two closed versions," one of which had response options emphasizing the low end of the sex partner distribution, the other emphasizing the high end. Dykema and Schaeffer (n.d.), reanalyzing the original study by Blair et al. plus additional experiments, concluded that "while closed questions result in higher reports of occurrence, the means among those engaging in the behaviors are usually greater with open questions." They attribute the difference in means to three factors: the composition of the sample, which is affected by whether a filter question is used; more frequent reporting of high frequencies with open questions; and whether those who report never engaging in the behavior are included in the analysis. In fact, they find that "open questions produce higher estimates of means for nonthreatening as well as threatening behaviors, and do not always do so for threatening questions (p. 24)."

Our research appears to have uncovered another version of this effect on the reporting of socially undesirable feelings. In research exploring race of interviewer effects using real and virtual interviewers, Krysan and Couper (2003) found some cases where white respondents (for example) gave more negative responses to live interviewers than to virtual ones. In qualitative debriefings of respondents, some mentioned that talking to an interviewer gave them an opportunity to explain their choice of responses; in the virtual interviewer condition (as on the web; see Krysan & Couper, 2005), they could only pick one of the response options provided, without the opportunity to justify their choice. Building on this observation, Couper (2012) conducted a web-based experiment in the Netherlands, using a series of questions on attitudes towards immigrants. Half the respondents were given the closed-ended

items, while the other half were presented both the closed-ended responses and an open-ended text box in which they could (if they wished) offer an explanation for their choice of responses. Surprisingly, offering such an option was associated with significantly more *positive* views towards immigrants. One alternative explanation is that the added open question encourages deeper cognitive processing of the question (i.e., thinking of reasons for or against endorsing the statement), potentially leading to more moderate views. Clearly, this finding suggests more research is needed on the role that optional open questions may play in the response process.

It is common in web surveys to limit respondents to one of the available options. In paper surveys, no such restrictions can be made (see Couper, 2008), and it is not uncommon for respondents to avail themselves of the opportunity to add additional information. Similarly, in interviewer-administered surveys, respondents often qualify their answers, express uncertainty, and the like. Much of this information is ignored in the interviewer's entry of the responses into the computer or in the keying of paper questionnaires. Automatic recording of the verbatim responses makes Schuman's (1966) idea of the random probe much more feasible, both for substantive and for methodological purposes. Adding such probes in web surveys, as Behr and her colleagues (2012, 2013, 2014) have shown, is relatively easy. If responses to such follow-up questions are not required, this is unlikely to have a negative effect on survey response.

A final use of open-ended questions is the "anything else" question sometimes appended to a structured questionnaire or interview: "Is there anything else you would like to tell us?" or "Are there any other comments you would like to make?" This is a write-in with a large text box in self-completion surveys (whether paper or web), or an open-ended question where the interviewer is supposed to record the answer verbatim, in interviewer-administered surveys. Often such responses are ignored or – at best – briefly scanned for key concerns, but rarely systematically coded and analyzed. Such a question may help give voice to respondents and may in turn provide us with valuable information, provided we make use of the information contained in the responses. As Peter Lynn has suggested (personal communication), such questions could be used to determine whether their inclusion affects response rates or panel attrition, or whether they affect related matters, such as respondent satisfaction with the survey.

3 Technological Developments Facilitating the Use of Open-Ended Questions

We've already noted that recent technical developments are facilitating both the capture and analysis of open-ended responses. In paper-based surveys, interviewers were either expected to transcribe the respondent's verbatim answers to open

questions, or to record them for later transcription, coding, and/or analysis. With the advent of computer-assisted interviewing, early concerns about requiring interviewers to type such responses into the computer proved to be largely unfounded (see, e.g., Bernard, 1989; Catlin & Ingram, 1988), but the introduction of CAPI may have served to accelerate the decline in the use of open-ended questions.

Field interviewers were often required to carry tape recorders to record entire interviews for quality control or methodological research purposes. Despite the intrusiveness of these devices, respondent consent rates to recording interviews were relatively high (see, e.g., Dykema et al., 1997; see also McGonagle et al., 2015). However, the equipment presented logistical difficulties for interviewers, both during and after the interview. Administrative effort was associated with labeling the cassettes and mailing them to a central office. Confidentiality concerns were raised regarding the handling and storage of the physical media. Analog cassette tapes also presented a big hurdle for coding and analysis. Coders had to search through the tapes to find the relevant sections, or be forced to listen to the entire interview. While the logistical and administrative procedures were somewhat less onerous in telephone surveys (e.g., larger keyboards made typing of open-ended responses easier, centralization facilitated the handling of recording equipment and cassettes), analysis of the responses or coding of the recordings remained a burdensome activity, and one that was hard to do selectively.

The development of digital recording devices made the capture of open-ended responses – and indeed the verbatim responses to all questions in the survey – much easier. Almost all laptop or tablet computers have the built-in capability for digital recording, obviating the need to carry additional equipment. Further, such recording can easily be integrated into the computer assisted interviewing (CAI) software (e.g., Thissen et al., 2013). Indeed, this tool, now known as computer-assisted recorded interviewing (CARI; see Arceneaux, 2007; Hicks et al., 2010; Thissen, 2014; Thissen et al., 2013) is a standard feature of some CAI systems. This brings several benefits for capture: 1) no need for additional equipment (although some laptop microphones are not ideal for recording both interviewers and respondents; see Hansen et al., 2005); 2) the consent process can be automated as part of the survey instrument (recording is automatically activated upon consent); 3) selected parts of the survey (sections or individual items) can be recorded; 4) sound files can be encrypted and transmitted to the central office as part of the regular send/receive activities; and 5) sound files can be easily identified (e.g., by sample ID and question number), facilitating the task of finding particular questions to listen to, transcribe and/or code. Although much of the work using CARI has focused on evaluating interviewer performance, we believe the tool also has great promise for revealing what respondents are saying – and how they are saying it, in response to both to open-ended and (ostensibly) closed-ended questions.

In similar fashion, the increased use of web surveys (relative to paper) makes the task of capturing typed responses to open-ended questions easier and more amenable to analysis. While adding such questions to web surveys may increase the perceived burden on respondents, making such questions optional may reduce this effect. Further, using randomization (as envisioned by Schuman, 1966) could further mitigate any negative consequences. Giving respondents an option to voice their own opinions may even have positive consequences, although this is largely untested (as we discuss elsewhere). Analysis of these responses from web surveys is facilitated by the fact that they are already in digital form.

Turning to analysis, a number of recent developments have made the analysis of open-ended data a much more tractable task. Specifically, recent improvements to several software packages for qualitative analysis make them more useful for the analysis of responses to open-ended questions (see Hughes, 2011). Further information on developments in the area of computer assisted qualitative data analysis (CAQDAS) can be found at the website http://www.surrey.ac.uk/sociology/research/researchcentres/caqdas/support/analysingsurvey/.

While not replacing the role of the researcher in developing and identifying themes for coding, these packages facilitate the task of coding itself.

The rapid development of software tools to facilitate the coding and analysis of textual materials in social media – whether through text mining or text analytics or more straightforward sentiment analysis – is expanding the opportunity for researchers to make use of fully-automated or semi-automated processes for coding of open text (see, e.g., Shonlau & Couper, 2016; Klochikhin & Boyd-Graber, 2017). Generic text analytic software, such as the natural language toolkit for Python (www.nltk.org; see also Bird et al., 2009) further facilitates the task of analysis. While these tools have not yet been widely embraced by survey researchers, and further exploration and evaluation is needed, they offer great promise for making the analysis of open-ended question less costly and time-consuming.

Another area of promising development lies in software to convert recorded speech to text. While such speech recognition software might not be ready for the task of converting large numbers of short segments to text (most systems require extensive training to improve recognition for a single user), they can potentially assist in substantially reducing the burden of manual transcription that is necessary for computer-assisted analysis of qualitative data. Recent advances in speech recognition, along with the development of powerful software tools to facilitate coding of text, promise to change the cost and effort equation for dealing with responses to open-ended questions.

4 Discussion

Instead of simply forcing respondents to agree (or otherwise) with the statements we proffer, or pick one of the responses we provide, we can give them an opportunity to tell us what's on their mind with respect to the topic under discussion – whether by offering an explicit open-ended question or by capturing everything they say during the interview. Wenemark (2010) suggests that this may empower and motivate respondents, and O'Cathain and Thomas (2004) go further in suggesting that open questions may help redress the power imbalance between researchers and respondents. However, this in turn obliges us to *listen* to what they say or *read* what they write.

We live in the digital age, where textual responses are readily analyzable using powerful text-analytic software, and where digital recordings of oral responses are increasingly amenable to automatic transcription. The cost of capturing this additional information has been dramatically reduced, and the ease with which it can be coded and analyzed has greatly increased. Yet we still seem to be operating as if paying attention to what respondents say – and the way they say it – is too costly and time-consuming for quantitative study.

The primary barriers to including open-ended responses in questionnaires or capturing verbatim responses relate to 1) concerns about lengthening the interview, 2) the risk of digression, 3) relying on interviewers faithfully recording the information, and 4) the cost of transcribing, coding and analyzing the resulting data. We address each of these objections briefly in turn.

The first two concerns are related. By encouraging respondents to provide open-ended responses, it is believed that interview length is increased and that "bad" respondent behavior is encouraged. Similarly, if interviewers are seen writing down everything that respondents say, this may encourage digression. While legitimate, these concerns are often taken to the extreme, leading to an avoidance of *any* open questions. By capturing responses unobtrusively, we reduce the risk of digression, and need to rely less on interviewers to record the responses as accurately as possible. Having interviewers paraphrase the respondents' answers to open-ended questions may still be valuable, but this could easily be supplemented with the actual words used by the respondents. Giving respondents an opportunity to voice their own views in their own words on key topics covered in the survey may well increase respondent engagement in the interview. This may be especially valuable in panel surveys, where cooperation in later waves is an important consideration.

The costs of processing and analyzing the open-ended responses remain a key concern. Recent software developments have made this a less-onerous undertaking, but it still requires effort. However, with digital recording, analysis can be done selectively, focusing on key questions identified prior to the start of data collec-

tion (e.g., those subject to random probes) or identified during data collection (e.g., by using paradata analysis of keystrokes to identify potentially problematic items). Further, selected subsets of interviews can be analyzed, potentially focusing on key subgroups of interest, such as those who provided a particular type of response or those who gave an indication of having difficulty with the question (again, as revealed through paradata; see Couper et al., 1997; Couper & Kreuter, 2011). In other words, technology has made it much easier to identify selected segments of an interview, and to identify subsets of interviews, questions, or respondents for more intensive analysis, reducing the effort and expense of such work.

As we have said earlier, we are not advocating a return to the days of unstructured interviews. Rather, we are arguing for the judicious use of open-ended questions to support the methodological goals outlined earlier. The verbatim responses we get to closed-ended questions, long ignored by survey researchers, may open up whole new areas of important methodological inquiry, providing valuable insights into the meaning and quality of the information respondents are providing as well as their motivation (or lack of it) for doing so.

We believe that the time has come to give greater voice to respondents in standardized surveys – to give them an opportunity, within the constraints of a structured interview, to express their views on the topics addressed in the survey in their own words. This is relevant to both interviewer-administered and self-administered surveys. Opening up the standardized survey in this way can be of benefit both to respondents (giving then a greater sense of engagement in the interaction) and to researchers (giving us more richly textured data on the topics we are studying and providing methodological insights into the process itself). Technological developments have facilitated this change, but inertia has inhibited us from using them to achieve these goals.

References

Arceneaux, T. A. (2007). Evaluating the Computer Audio-Recorded Interviewing (CARI) Household Wellness Study (HWS) Field Test. *Proceedings of the Survey Research Methods Section, American Statistical Association*, 2811-2818.

Bates, N. A., Dahlhamer, J., & Singer, E. (2008). Privacy Concerns, Too Busy, or Just Not Interested: Using Doorstep Concerns to Predict Survey Nonresponse. *Journal of Official Statistics,* 24(4), 591-612.

Behr, D., Kaczmirek, L., Bandilla, W., & Braun, M. (2011). Asking Probing Questions in Web Surveys: Which Factors Have an Impact on the Quality of Responses? Mannheim: GESIS working paper.

Behr, D., Braun, M., Kaczmirek, L., & Bandilla, W. (2012). Asking Probing Questions in Web Surveys: Which Factors have an Impact on the Quality of Responses. *Social Science Computer Review,* 30(4), 487-498.

Behr, D., Kaczmirek, L., Bandilla, W., & Braun, M. (2013). Testing the Validity of Gender Ideology Items by Implementing Probing Questions in Web Surveys. *Field Methods,* 25(2), 124-141.

Behr, D., Bandilla, W., Kaczmirek, L., & Braun, M. (2014). Cognitive Probes in Web Surveys: On the Effect of Different Text Box Size and Probing Exposure on Response Quality. *Social Science Computer Review,* 32(4), 524-533

Benkí, J. R., Broome, J., Conrad, F. G., Groves, G. R., & Kreuter, F. (2011). Effects of Speech Rate, Pitch, and Pausing on Survey Participation Decisions. Paper presented at the annual conference of the American Association for Public Opinion Research, Phoenix, May.

Belson, W. A. (1981). *The Design and Understanding of Survey Questions*. Aldershot, UK: Gower.

Bernard, C. (1989). Survey Data Collection Using Laptop Computers. Paris: INSEE (Report No. 01/C520).

Bird, S., Klein, E., & Loper, E. (2009). *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. Boston: O'Reilly Media.

Blair, E., Sudman, S., Bradburn, N. M., & Stocking, C. (1977). How to Ask Questions about Drinking and Sex: Response Effects in Measuring Consumer Behavior. *Journal of Marketing Research,* 14, 316-321.

Campanelli, P., Sturgis, P, & Purdon, S. (1997). Can You Hear Me Knocking: An Investigation into the Impact of Interviewers on Survey Response. London: Social and Community Planning Research.

Cannell, C. F. & Oksenberg, L. (1988). Observation of Behavior in Telephone Interviews. In R. M. Groves, P. P. Biemer, L. E. Lyberg, J. T. Massey, W. L. Nicholls II, & J. Waksberg (eds.), *Telephone Survey Methodology* (pp. 475-495). New York: Wiley.

Catlin, G., & Ingram, S. (1988). The Effects of CATI on Costs and Data Quality: A Comparison of CATI and Paper Methods in Centralized Interviewing. In R. M. Groves, P. P. Biemer, L. E. Lyberg, J. T. Massey, W. L. Nicholls II, & J. Waksberg (eds.), *Telephone Survey Methodology* (pp. 437-450). New York: Wiley.

Conrad, F. G., Broome, J. S., Benki, J. R., Kreuter, F., Groves, R. M., Vannette, D., McClain, C., et al. (2013). Interviewer Speech and the Success of Survey Invitations. *Journal of the Royal Statistical Society, Series A,* 176(1), 191-210.

Couper, M. P. (1997). Survey Introductions and Data Quality. *Public Opinion Quarterly,* 61(2), 317-338.

Couper, M. P. (2008). Technology and the Survey Interview/Questionnaire. In M. F. Schober & F. G. Conrad (eds.), *Envisioning the Survey Interview of the Future* (pp. 58-76). New York: Wiley.

Couper, M. P. (2012). Reducing the Threat of Sensitive Questions in Online Surveys. Paper under preparation for presentation at the General Online Research Conference, Mannheim, March.

Couper, M. P., & Groves, R. M. (2002). Introductory Interactions in Telephone Surveys and Nonresponse. In D. W. Maynard, H. Houtkoop-Steenstra, N. C. Schaeffer, & J. van der Zouwen (eds.), *Standardization and Tacit Knowledge; Interaction and Practice in the Survey Interview* (pp. 161-177). New York: Wiley.

Couper, M. P., Hansen, S. E., & Sadosky, S. A.. (1997). Evaluating Interviewer Performance in a CAPI Survey. In L. E. Lyberg et al. (eds.), *Survey Measurement and Process Quality* (pp. 267-285). New York: Wiley.

Couper, M. P., Kennedy, C., Conrad, F. G., & Tourangeau, R. (2010). Designing Input Fields for Non-Narrative Open-Ended Responses in Web Surveys. *Journal of Official Statistics,* 27(1), 65-85.

Couper, M. P., & Singer, E. (2013). Informed Consent for Web Paradata Use. *Survey Research Methods,* 7(1), 57-67.

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, 255-275.

Couper, M. P., Singer, E., Conrad, F. G., & Groves, R. M. (2010). Experimental Studies of Disclosure Risk, Disclosure Harm, Topic Sensitivity, and Survey Participation. *Journal of Official Statistics,* 26(2), 287-300.

DeMaio, T. J. (1983). *Approaches to Developing Questionnaires, Statistical Policy Working Paper 10*, Subcommittee on Questionnaire Design, Federal Committee on Statistical Methodology. Washington, DC: Office of Management and Budget.

DeMaio, T. J., & Rothgeb, J. (1996). Cognitive Interviewing Techniques: In the Lab and in the Field. In Norbert Schwarz and Seymour Sudman (eds.), *Answering Questions: Methodology for Determining Cognitive and Communicative Processes in Survey Research (*pp. 177-195). San Francisco: Jossey-Bass.

Dykema, J., Lepkowski, J. M., & Blixt, S. (1997). The Effect of Interviewer and Respondent Behavior on Data Quality: Analysis of Interaction Coding in a Validation Study. In P. P. Biemer, R. M. Groves, L. E. Lyberg, N. A. Mathiowetz, & S. Sudman (eds.), *Measurement Errors in Surveys* (pp. 287-310*)*. New York: Wiley.

Dykema, J., & Schaeffer, N. C. (n.d.). Effects of Open, Closed, and Filter Questions on Measuring the Occurrence and Frequency of Behaviors. Unpublished ms., University of Wisconsin, Madison.

Fowler, F. J., & Cannell, C. F. (1996). Using Behavioral Coding to Identify Cognitive Problems with Survey Questions. In N. Schwarz & S. Sudman (eds.), *Answering Questions: Methodology for Determining Cognitive and Communicative Processes in Survey Research* (pp. 15-36). San Francisco: Jossey-Bass.

Fuchs, M. (2009). Asking for Numbers and Quantities: Visual Design Effects in Paper and Pencil Surveys. *International Journal of Opinion Research,* 21(1), 65-84.

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

Geer, J. G. (1991). Do Open-Ended Questions Measure Salient Issues? *Public Opinion Quarterly,* 55(3), 360-370.

Groves, R. M., & Couper, M. P. (1996). Contact-Level Influences on Cooperation in Face-to-Face Surveys. *Journal of Official Statistics,* 12(1), 63-83.

Groves, R. M., & McGonagle, K. A. (2001). A Theory-Guided Interviewer Training Protocol Regarding Survey Participation. *Journal of Official Statistics,* 17(2), 249-266.

Hansen, S. E., Krysan, M, & Couper, M. P. (2005). Sound Bytes: Capturing Audio in Survey Interviews. Poster presented at the annual meeting of the American Association for Public Opinion Research, Miami Beach, May.

Hicks, W. D., Edwards, B., Tourangeau, K., McBride, B., Harris-Kojetin, L. D., & Moss, A. J. (2010). Using CARI Tools to Understand Measurement Error. *Public Opinion Quarterly,* 74(5), 985-1003.

Hughes, G. (2011). Using Qualitative Software to Analyse Open-Ended Survey Questions: A Guide for Quantitative Analysts. Paper presented at the European Survey Research Association Conference, Lausanne, July.

Hess, J. C., & Singer, E. (1995). The Role of Respondent Debriefing Questions in Questionnaire Development. *Statistical Research Division Working Papers in Survey Methodology (#95-18).* U.S. Census Bureau. Available online at http://www.census.gov/srd/papers/pdf/sm9518.pdf.

Israel, G. D. (2010). The Effects of Answer Space Size on Responses to Open-Ended Questions in Mail Surveys. *Journal of Official Statistics,* 26(2), 271-285.

Klochikhin, E., & Boyd-Graber, J. (2017). Text Analysis. In I. Foster, R. Ghani, R. S. Jarmin, F. Kreuter, & J. Lane (eds.), *Big Data and Social Science: A Practical Guide to Methods and Tools* (pp. 187-214). Boca Taron, FL: CRC Press.

Krysan, M., & Couper, M. P. (2003). Race in the Live and Virtual Interview: Racial Deference, Social Desirability, and Activation Effects in Attitude Surveys. *Social Psychology Quarterly,* 66(4), 364-383.

Krysan, M., & Couper, M. P. (2005). Race-of-Interviewer Effects: What Happens on the Web? *International Journal of Internet Science,* 1(1), 5-16.

Lazarsfeld, P. F. (1935). The Art of Asking Why. *National Marketing Review,* 1, 32-43.

Lazarsfeld, P. F. (1944). The Controversy over Detailed Interviews: An Offer for Negotiation. *Public Opinion Quarterly,* 8, 38-60.

Martin, E. (2004). Vignettes and Respondent Debriefing for Questionnaire Design and Evaluation. In S. Presser, J. Rothgeb, M. P. Couper, J. T. Lessler, E. A. Martin, J. Martin, & E. Singer (eds.), *Methods for Testing and Evaluating Survey Questionnaires* (pp. 149-171). New York: Wiley.

Mason, R., Carlson, J. E., & Tourangeau, R. (1994). Contrast Effects and Subtraction in Part-Whole Questions. *Public Opinion Quarterly,* 58(4), 569-578.

Maynard, D., Freese, J., & Schaeffer, N. C. (2010). Calling for Participation: Requests, Blocking Moves, and Rational (Inter)Action in Survey Introductions. *American Sociological Review,* 75(5), 795-814.

Maynard, D., Houtkoop-Steenstra, H., Schaeffer, N. C., & van der Zouwen, J. (eds.). (2002). *Standardization and Tacit Knowledge; Interaction and Practice in the Survey Interview.* New York: Wiley,

McGonagle, K. A., Brown, C, & Schoeni, R. F. (2015). The Effects of Respondents' Consent to Be Recorded on Interview Length and Data Quality in a National Panel Study. *Field Methods,* 27(4), 373-390.

Morton-Williams, J. (1993). *Interviewer Approaches.* Aldershot: Dartmouth.

Morton-Williams, J., & Young, P. (1987). Obtaining the Survey Interview – An Analysis of Tape-Recorded Doorstep Introductions. *Journal of the Market Research Society,* 29(1), 35-54.

O'Cathain, A., & Thomas, K. J. (2004). 'Any Other Comments?' Open Questions on Questionnaires – A Bane or a Bonus to Research? *BMC Medical Research Methodology,* 4(25).

Oksenberg, L., Cannell, C. F., & Kalton, G. (1991). New Strategies for Pretesting Survey Questions. *Journal of Official Statistics*, 7, 349-356.

Porst, R. & von Briel, C. (1995). Wären Sie vielleicht bereit, sich gegebenenfalls noch einmal befragen zu lassen? Oder: Gründe für die Teilnahme an Panelbefragungen. ZUMA-Arbeitsbericht, Nr. 95/04, Mannheim, Germany.

Presser, S., Rothgeb, J., Couper, M. P., Lessler, J. T., Martin, E., Martin, J., & Singer, E. (eds.). (2004). *Methods for Testing and Evaluating Survey Questionnaires.* New York: Wiley.

Sakshaug, J. W., Couper, M. P., & Ofstedal, M. B.. (2010). Characteristics of Physical Measurement Consent in a Population-Based Survey of Older Adults. *Medical Care,* 48(1), 64-71.

Sakshaug, J. W., Couper, M. P., Ofstedal, M. B., & Weir, D. (2012). Linking Survey and Administrative Records: Mechanisms of Consent. *Sociological Methods and Research,* 41(4), 535-569.

Schaeffer, N. C., & Maynard, D. W.. (1996). From Paradign to Prototype and Back Again: interactive Aspects of Cognitive Processing in Standardized Survey Interviews. In N. Schwarz & S. Sudman (eds.), *Answering Questions: Methodology for Determining Cognitive and Communicative Processes in Survey Research* (pp. 65-88). San Francisco: Jossey-Bass.

Schonlau, M., & Couper, M. P. (2016). Semi-Automated Categorization of Open-Ended Questions. *Survey Research Methods*, 10(2), 143-152.

Schuman, H. (1966). The Random Probe: A Technique for Evaluating the Validity of Closed Questions. *American Sociological Review,* 31(2), 218-222.

Schuman, H. (2008). *Method and Meaning in Polls and Surveys.* Cambridge, MA, Harvard University Press.

Schuman, H., & Presser, S. (1977). Question Wording as an Independent Variable in Survey Analysis. *Sociological Methods and Research,* 6, 151-170.

Schuman, H., & Presser, S. (1979). The Open and Closed Question. *American Sociological Review,* 44(5), 692-712.

Singer, E. (2003). Exploring the Meaning of Consent: Participation in Research and Beliefs about Risks and Benefits. *Journal of Official Statistics,* 19, 273-286.

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., & Couper, M. P. (2011). Ethical Considerations in Web Surveys. In M. Das, P. Ester, & L. Kaczmirek (eds.), *Social Research and the Internet* (pp. 133-162). New York: Taylor and Francis.

Singer, E., & Kulka, R. A. (2002). Paying Respondents for Survey Participation. In *Studies of Welfare Populations: Data Collection and Research Issues*, edited by M. Ver Ploeg, R. A. Moffitt, & C. F. Citro. Washington, D.C.: National Academy Press.

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

Sturgis, P., & Campanelli, P. (1998). The Scope for Reducing Refusals in Household Surveys: An Investigation Based on Transcripts of Tape-Recorded Doorstep Interactions. *Journal of the Market Research Society,* 40(2), 121-39.

Tan, L. (2011). An Introduction to the Contact History Instrument (CHI) for the Consumer Expenditure Survey. Washington, DC: Bureau Labor Statistics, Consumer Expenditure Survey Anthology, 2011, pp. 8-16, available at http://www.bls.gov/cex/anthology11/csxanth2.pdf

Thissen, M. R. (2014). Computer Audio-Recorded Interviewing as a Tool for Survey Research. *Social Science Computer Review*, 32(1), 90-104.

Thissen, M. R, Fisher, C., Barber, L., & Sattaluri, S. (2008). Computer Audio-Recorded Interviewing (CARI), A Tool for Monitoring Field Interviewers and Improving Field Data Collection. Proceedings of the International Methodology Symposium 2008, Statistics Canada, Gatineau, Canada.

Thissen, M. R., Park, H., & Nguyen, M. (2013). Computer Audio Recording: A Practical Technology for Improving Survey Quality. *Survey Practice*, 6(2).

Thissen, M. R., & Rodriguez, G. (2004). Recording Interview Sound Bites Through Blaise Instruments. Paper presented at the International Blaise User's Conference, Gatineau, Québec, Canada, September.

Tourangeau, R., Conrad, F. G., Couper, M. P., & Ye, C. (2014). The Effects of Providing Examples in Survey Questions. *Public Opinion Quarterly,* 78(1), 100-125.

Tourangeau, R., Sun, H., Conrad, F. G., & Couper, M. P. (2016). Examples in Open-Ended Survey Questions. *International Journal Public Opinion Research*, advance online access, DOI: 10.1093/ijpor/edw015.

Tourangeau, R., & Smith, T. W. (1996). Asking Sensitive Questions: The Impact of Data Collection Mode, Question Format, and Question Content. *Public Opinion Quarterly,* 60(2), 275-304.

Wenemark, M. (2010). *The Respondent's Perspective in Health-Related Surveys.* Linköping University Medical Dissertations No. 1193.

West, B. T. (2013). An Examination of the Quality and Utility of Interviewer Observations in the National Survey of Family Growth. *Journal of the Royal Statistical Society, Series A,* 176(1), 211-225.

West, B. T., & Kreuter, F. (2013). Factors Affecting the Accuracy of Interviewer Observations: Evidence from the National Survey of Family Growth (NSFG). *Public Opinion Quarterly,* 77(2), 522-548.

West, B. T., & Kreuter, F. (2015). A Practical Technique for Improving the Accuracy of Interviewer Observations of Respondent Characteristics. *Field Methods,* 27(2), 144-162.

Yan, T., Curtin, R., & Jans, M. (2010). Trends in Income Nonresponse over Two Decades. *Journal of Official Statistics,* 26(1), 145-164.

methods, data, analyses | Vol. 11(2), 2017, pp. 135-162 DOI: 10.12758/mda.2017.02

Analyzing Survey Characteristics, Participation, and Evaluation Across 186 Surveys in an Online Opt-In Panel in Spain

Melanie Revilla
RECSM-Universitat Pompeu Fabra

Abstract

Survey designers often ask about the best length for their questionnaires and the best format for their questions. Much research has already addressed these issues. However, the answers to these questions may vary with the population of interest, the mode of data collection used, and other factors.

The goal of this paper is twofold:

1. To give an overview of the present situation in opt-in online panels, in terms of survey characteristics, participation, and evaluation, by reviewing 186 surveys managed by the panel company Netquest in Spain in 2016. This will be useful to determine areas where further research needs to focus.
2. To study how key characteristics of questionnaires impact survey evaluation and levels of survey break-off. This will allow us to highlight the characteristics that best reduce break-off and improve respondents' survey evaluation.

Based on these results, we will propose practical recommendations for future survey design within the framework of opt-in online panels.

Keywords: opt-in online panels, questionnaire characteristics, participation, survey evaluation, mobile devices

© The Author(s) 2017. This is an Open Access article distributed under the terms of the Creative Commons Attribution 3.0 License. Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.

1 Introduction

Research is increasingly relying on survey data, and thus on individuals' willingness to participate in surveys and provide quality responses. Designing and implementing a survey requires numerous decisions, all of which may affect respondents' willingness to participate in a given survey or accept future invitations, as well as the overall quality of the data obtained. In order to achieve high participation and high overall data quality, when designing new surveys, frequent questions include: What is the advised length of a questionnaire?[1] How should questions be formatted? What can be done to limit break-off?

The goal of this paper is twofold:

1. By reviewing 186 surveys run by the panel company Netquest in Spain in 2016, the paper aims to provide an overview of the current situation in opt-in online panels, in terms of: survey characteristics (e.g., target populations, quotas, survey content, including topic, question formats, estimated survey length, and incentives), participation (i.e. the number of panelists invited, the number that began the survey, and the numbers that screened out, broke off, or completed the entire survey), and evaluation of the survey itself (each survey included a final question allowing respondents to evaluate the survey they just finished, on a scale from "1-survey very badly done" to "5-survey very well done").

As Netquest provides data to all kinds of clients, agencies and researchers, we expect this overview to allow us to identify which target populations, survey topics, and question formats are most commonly used over a large range of research. This can be useful for at least two reasons: first, it helps us determine areas where further research is needed; second, it helps us identify areas where there are large disparities between the knowledge found in literature and what is done in practice. For instance, if the literature provides clear evidence that a specific question format performs worse than another, but we observe that the less efficient format is used more often in practice, we know where to channel our efforts when transferring knowledge from the academic world to the real practice of online surveying.

---

1 This is, for instance, what the author claims in this post: https://www.surveygizmo.com/survey-blog/how-long-can-a-survey-be/

Acknowledgements

I would like to acknowledge Netquest for providing me with the necessary data, and thank Gerardo Ortiz in particular for his huge support in getting the information needed, as well as Carlos Ochoa for his useful advice throughout the process. I also thank two anonymous reviewers and the Editor for their useful comments on a previous draft of this manuscript.

Direct correspondence to

Melanie Revilla, RECSM-Universitat Pompeu Fabra, Mercè Rodoreda 24.406, Ramón Trías Fargas, 25-27, 08005 Barcelona, Spain
E-mail: melanie.revilla@hotmail.fr or melanie.revilla@upf.edu

This overview also presents the overall level of participation and break-off, as well as respondents' average evaluation of surveys. These aspects may be more specific to the panel studied, as practical decisions concerning incentives or whether to announce survey length could affect these variables.

2. The paper also studies how break-off levels and survey evaluation are related to key characteristics of the questionnaires: topic, question type, and estimated survey length.

This second part seeks to identify whether survey-design decisions affect the break-off rate and participants' evaluation of the survey, and, if so, which decisions matter more. There is a tremendous unmet demand from online opt-in panels for survey-design guidelines. Based on our results, we will make some practical recommendations for future survey design within the framework of opt-in online panels.

The remainder of this paper is organized as follows: Section 2 offers background information, Section 3 provides information about the methodology, data and analyses conducted, Section 4 presents the main results, and Section 5 concludes.

2 Background

Much research has been done on survey characteristics, and it has chiefly focused on the effects of the topic, question format, survey length, and incentives on survey participation (mainly in terms of response rates) and on other aspects of data quality, such as break-off rates and survey evaluation (see Schuman & Presser, 1981; Sudman & Bradburn, 1982; Oppenheim, 1992; Tourangeau, Rips, & Rasinski, 2000; Brace, 2004; Saris & Gallhofer, 2014).

Most of these studies investigated face-to-face, telephone or postal mail surveys. However, in the past 15 to 20 years, web surveys have gained increasing traction. This new data collection mode differs at several levels from more traditional modes (de Leeuw, 2005): for instance, web surveys are computer-assisted and self-completed, and the stimulus is usually visual. Because it is easier to close a tab than to ask an interviewer to leave your home half-way through the questionnaire, it is expected that more respondents will break off online than during face-to-face surveys. Besides, with web surveys, respondents cannot turn to the interviewer if they have difficulty understanding or experience technical problems (e.g., if the webpage does not load). Thus, question layout and formulation could be even more important than in a face-to-face survey. In addition, different recommendations were needed about how to design these web surveys. This generated a lot of new research (e.g., Couper, 2000; Couper, Traugott, & Lamias, 2001; Dillman, 2000; Dillman & Bowker, 2001; Lozar Manfreda, Batagelj, & Vehovar, 2002; Marcus,

et al., 2007; Couper, 2008; Galesic & Bosnjak, 2009; Bethlehem & Biffignandi, 2011; Tourangeau, Conrad, & Couper, 2013), from which lists of recommendations have been extracted: see, for instance, Parsons' paper (2007) on web survey best practices.

Nevertheless, as Couper and Miller (2008) have pointed out, web surveys can be extremely different, and what applies in one case does not necessarily hold in another. One crucial distinction, in particular when studying survey participation, is the difference between one-time surveys and surveys done within the framework of online panels, in which the same group of people who agreed to participate in surveys are regularly contacted by the same online panel company to complete questionnaires, usually in exchange for money or gifts, as this helps increase collaboration (Göritz, 2006). By nature, panels need to retain respondents. Consequently, survey experience is more important for panels, as it can affect future participation. Cape (2012, p.6) stresses the need to find better ways to motivate online panelists, and recommends moving them toward "intrinsic motivation" to keep them active: "[Online panel companies] have a finite resource, which costs money to build and develop, and the industry as a whole is forcing down revenues per interview. The more we can do to motivate our panelists, the easier they will be to recruit and retain."

In addition, within online panels, it is important to differentiate between probability-based panels and opt-in or access panels (Callegaro, Lozar Manfreda, & Vehovar, 2015, chapter 5.2.). In probability-based panels, a random sample is drawn from the population and the selected units are contacted and invited to participate in the panel. Units who do not have Internet access are usually provided with it. On the other hand, in opt-in or access panels, individuals volunteer to participate. If they do not have Internet access, they cannot be part of the panel. This raises the issues of how representative for the target population different online panels truly are, and of how to deal with samples whose panelist profile may differ from that of the target population (e.g., using weighing; Callegaro, Lozar Manfreda, & Vehovar, 2015, 5.2), in terms not just of socio-demographics, but also of attitudinal variables. Because people volunteer to participate in opt-in panels, there is also a risk of professional respondents, that is respondents who frequently participate in surveys and are mainly doing so for incentives (Mathijsse, de Leeuw, & Hox, 2015). This could affect data quality in a number of ways (e.g., if these respondents are speeding through the questionnaire; see, e.g., Zhang & Conrad, 2014).

Probability-based and opt-in panels usually differ in several additional respects, including: a) the frequency of contact with panelists (more frequent in opt-in panels), b) panel management (sending similar surveys to all panelists versus sending completely different surveys to subgroups of panelists depending on their profile), c) the kind of survey sent (mainly academic versus mainly commercial), d) their goals (to represent the general population or to cover very specific target pop-

ulations needed by the client), etc. On the one hand, all this suggests that different recommendations might be needed for these opt-in panels. On the other hand, in the last few decades, this form of collecting web survey data has become common, in particular in market research, but also in other areas, such as social and political science. According to the AAPOR Standards Committee (2010), the majority of online research is based on non-probability panels. Thus, we believe that they require special attention.

Nevertheless, the opt-in online panels have not been studied much yet, although in a few cases methodological research is moving in that direction. For instance, Stenbjerre and Laugesen (2005) offer a summary of five years' worth of lessons learned while working with the Zapera online access panels in the Nordic region (Denmark, Sweden, Norway, Finland and Estonia). They approach the issue from several different directions, including recruitment, participation and incentives. More recently, Cape and Phillips (2015) examined the effects of questionnaire length on data quality for an opt-in online panel. Two books (Callegaro et al., 2014; Callegaro, Lozar Manfreda, & Vehovar, 2015) that include several chapters focused on non-probability-based online panels were recently published; they cover issues such as panelists' motivation for joining non-probability online panels, speeding, and professional respondents. These books include a summary of the results of the NOPVO study, the first large-scale commercial study to compare different non-probability panels in the Netherlands. This study a) made an inventory of all online panels in the Netherlands and b) compared the results across 19 online panels that conducted a similar survey (see also Vonk, Van Ossenbruggen, & Willems, 2006). Later, similar comparisons were conducted in the USA (Walker, Pettit, & Rubinson, 2009) and in Canada (Chan & Ambrose, 2011). All these studies revealed significant differences across online opt-in panels.

In addition, web surveys are increasingly completed on tablets and smartphones (Callegaro, 2010; De Bruijne & Wijnant, 2014; Revilla et al., 2016), which differ from traditional PCs in important ways: they are smaller, have touch-screens, are portable, etc. (Peytchev & Hill, 2010; Wells, Bailey, & Link, 2013). Thus, different recommendations may be needed when these mobile devices are used by at least some respondents.

In this paper, we focus on opt-in online panels, as these are increasingly common, differ from other means of collecting data on many levels, and have not yet received much academic attention. When available, our analyses also consider information on the devices panelists used to complete their surveys.

3 Methodology

3.1 Data: All Surveys Programmed by Netquest and Answered by Netquest Panelists in Spain Over a Period of About 6 Months

Our data comes from the Netquest online fieldwork company (www.netquest.com). Netquest has opt-in online panels in several countries since 2006. Netquest sends panelists survey invitations via email, using a list of individuals who have agreed to receive emails after answering a short satisfaction survey on a website belonging to one of the company's many collaborators. For each survey completed, panelists are rewarded with points, based on the estimated length of the questionnaire. These points can be exchanged for gifts. The company has panels in 23 countries. In this study, we focus on Spain, where the current panel counts 117,001 active panelists.

Our first goal was to get a good overview of what is the current situation in the Netquest panel in Spain in terms of survey characteristics, participation, and evaluation. To do this, we considered all surveys implemented by Netquest in Spain for a period of about six months (from mid-February 2016 to beginning of August 2016). We were interested in surveys that (1) were programmed by Netquest, so that we could have access to all necessary information, and (2) were sent to Netquest panelists (not to external databases provided by clients). A total of 216 surveys corresponded to this target of interest. However, we excluded 30 surveys, because of different reasons:

- Two were sent to Netquest panelists but were completed by their children; thus, they studied a different population.
- One study wanted only 15 interviews; as this was really a special case, we preferred to discard it.
- In 27 surveys, metadata was not properly collected, so we could not access necessary information on survey evaluation, devices used, etc.

In the end, 186 surveys were included in the database that we created by coding the characteristics of each individual survey. Five surveys were missing information on some aspects of interest, but the absences were minimal, so we kept them.

3.2 Aspects considered for the overview

We were interested in different aspects of each survey.

First, who is the target population? Besides the text description of each target population, we coded the following aspects:

- General or specific target population. We counted the target as the "general population" even when age limits were defined if these ages were between 16 (or 18) and 65 (or more). We also counted surveys targeting the general Internet population as the "general population". Thus, this is a quite broad definition of the "general population."
- Populations including only one gender: Surveys targeting only men or only women.
- Populations including limits on age, besides the 16+ or 18+.
- More than one target populations: For instance, surveys asking for 500 male respondents from 25 to 50, and 500 respondents who used product X at least once a week.
- We also research quotas used (if any).

***Second, what are the questionnaires' characteristics*?** In this case, we used the questionnaires to determine the main topics of the surveys as well as the main question formats used:

- Grids (also called "battery"), in which several items are presented together in a matrix format. Many studies contrast grid questions with item-by-item formats (see Tourangeau, Conrad, & Couper 2013, p. 72-76 for a summary). Even if results from the literature are mixed, many practitioners argue against the use of grids (e.g., Poynter, 2001 or Dillman, Smyth, & Christian, 2009), in particular when there are smartphone respondents (Lorch & Mitchell, 2014).
- Open-ended questions, in which respondents have to type in text as an answer. While closed questions have the advantage of being easier to analyze (they do not need to be coded) and may require less effort from respondents, open questions allow more elaborate answers. However, there are concerns that these questions might not provide all the information expected, particularly when respondents use mobile devices (e.g., Lambert & Miller, 2015).
- Multiple-response questions, in which the respondents can/must select all options they want or all options that apply. The instructions do not always explicitly state that respondents must "check all that apply," but there is no limit on the number of items the respondents can select. Previous research has usually recommended avoiding multiple-response questions and using "forced-choice" formats (e.g., asking to say yes or no for each item) instead (e.g., Smyth et al., 2006).
- Sliders, in which respondents have to position themselves on a sort of line. Again, results from previous research usually suggest that simpler alternative scales like radio buttons perform better (e.g., Funke, Reips, & Thomas, 2011).
- Dropdowns menus, in which respondents must click to make the menu appear and then select the most adequate option. Once more, there is some evidence against the use of drop-down menus (e.g., Healey, 2007).

- Ordering questions, in which respondents must rank different items from a list. Concerns have been raised about the measurement properties of ranking versus rating tasks (see, e.g., Ovadia, 2004).

We also coded if the questionnaires included some "agree/disagree" questions, that is questions asking explicitly if respondents agree or disagree with certain statements. Indeed, previous research suggests that this format creates a higher cognitive burden (Fowler, 1995) and acquiescence bias (Krosnick, 1991), as well as lower measurement quality (Saris et al., 2010).

Furthermore, we checked surveys to see if videos were present, as this can lead to more technical problems (i.e., panelists having troubles to viewing videos, in particular on smartphones).

We should note that, in general, in questionnaires programmed by Netquest, respondents cannot continue to the next question without providing an answer to the current question. Nevertheless, because of the presence of filter questions, all respondents in a given survey do not always get the same questions. We considered a format to be present if the highest proportion of respondents within a given survey got at least one question in that format. Thus, if 80% of respondents did not get a slider, and 20% did, we coded the survey as having "0 slider".

Finally, in this section we also consider the estimated length of the survey, which Netquest uses to determine the incentive respondents receive for each survey. The question of whether an ideal questionnaire length exists was already discussed in 1981 by Herzog and Bachman (p. 549). While some researchers "are convinced that survey instruments have a maximum length beyond which there is an increasing probability of premature termination, random responding, or other behavior patterns which result in data of lower quality," others "argue that a survey can be quite long without serious loss of respondents or deterioration in the quality of the responses." These authors found a tendency of somewhat lower quality answers toward the end of long questionnaires.

Third, what did respondents receive in exchange for their participation? Here, we focus on the incentives participants received for completing the entire survey. Incentives are in the form of points, which can be exchanged for gifts.

Fourth, what happened during the fieldwork? Panelists are invited to participate in a given survey. In Netquest's case, profiling information (i.e. information on different aspects of the panelists' lives, in particular behaviors and buying habits, which the panel organization has already collected and stored) is used, when available, to invite individuals who are expected to fit the target population. Once they receive the invitation, panelists can decide to start the survey (we will refer to this case as "started") or not. In the case of Netquest, panelists normally do not get any information about the survey in the invitation, so the decision to participate cannot

be linked to the survey's characteristics. Once a panelist starts the survey, different scenarios are possible:

- The panelist does not fit the population of interest or does not fit the set quotas (some quotas are already full). Thus he/she will be excluded from the survey and redirected to a profiling module. We refer to this case as "screened out."
- The panelist decides by him/herself to abandon the survey. This can occur at any moment after the panelist has started. We refer to this as "break-off."
- The panelist reaches the survey's final question. We refer to this as "complete."

We report the number of invitations, surveys started, panelists screened out, break-offs, and surveys completed across all surveys. From these numbers, we also calculated the following:

- Participation Rate = *(number started / number invited)*100*
- Screen-out Rate = *(number screened out /number started)*100*
- Break-off Rate = *[number of break-offs / (number of completes + number of break-offs)]*100*

Furthermore, for the panelists who completed the whole survey, we also considered the type of device (PC, tablet or smartphone) they used and the number of sessions in which they completed the survey (recorded automatically).

Fifth, what was the average evaluation of each survey? At the end of each survey, we added a question asking respondents to evaluate the survey, from (1) very badly done to (5) very well done. We considered the average across all respondents (PC, tablet, and smartphone) in each survey, as well as the average for PC-only and Smartphone-only respondents. We did not consider tablets separately, as they were used in a low number of cases.

3.3 How the Break-off Rate and Survey Evaluation Relate to key Questionnaire Characteristics

After our overview, we examined the relationships between some of the aspects considered. We do not study the decision to start the survey, as this is cannot be related to survey characteristics (no information is provided before the survey starts), nor did we study the screen-out rate, as this depends on the population of interest and the quotas required. On the contrary, the break-off rate is determined by panelists' decisions, and can be affected by survey characteristics such as the topic, the question format, and the estimated length. Galesic (2006) found that break-off is related to low interest (which can be linked to the topic) and higher reported burden (which can be linked to question format). Yan et al. (2010) consider the link between break-off and the interaction among the task duration announced, the real number of questions, and the presence of a progress indicator. In the book

by Tourangeau, Conrad, and Couper (2013), chapter 3.6 is dedicated to the "Factors Affecting Break-offs in Web Surveys". However, most of this research is not primarily or not at all focused on the case of online opt-in panels.

We also consider the impact of survey characteristics on respondents' evaluation of the survey. Indeed, if respondents do not like the survey, they may abandon it. In addition, if they do complete it, their satisfaction with the survey experience is expected to be lower. In this case, we expect both their probability of accepting their next survey invitation and the quality of their answers in the current survey to be reduced. This idea is supported by Cape and Phillips (2015), who found that longer surveys in an online opt-in panel do not lead to increased break-off rates, but are correlated with people speeding up during the survey, with higher satisficing, and thus lower data quality. Therefore, it is also important to study respondents' opinions of the survey.

We should note, however, that if the survey evaluation can be determined by the general characteristics of the survey, the break-off rate can only be affected by the characteristics of the questions prior to the break-off point. Nevertheless, we are not able to take this into account in our analyses, as we only possess information aggregated at the survey level. This is a key limit to these analyses. We are also limited by the fact that we possess information about the survey evaluation from the panelists who finished the survey, but not from those who did broke off or were screened out.

4 Main Results

4.1 Overview of the Current Situation in an Online Opt-in Panel

4.1.1 What is the Target Population for these Surveys?

First, we looked at the target population of the 186 surveys in our database. Table 1 presents a few examples of target population definitions. Table 2 gives the proportions of surveys that: have the general population as their target population; are limited to one gender; have age limits besides 16+ or 18+; and have more than one target of interest.

The examples in Table 1 give an idea of how specific the target populations can be in surveys run in online opt-in panels such as Netquest. It also shows how problematic getting a representative sample of such populations can be. For instance, individuals who need orthodontic work may not be aware of that fact. These hypothetical people would answer that they do not need an orthodontist, and could be screened out of surveys trying to target them. How can researchers

Table 1 Population of interest: a few examples

Definition of target population
- Men and women from 25-50 years old who play sports at least twice a week with an intensity of one hour of athletic activity and who bought detergent in the last two months.
- Population that has suffered from or is suffering from gout.
- 30-65 year-olds who consume oat, almond or rice drinks.
- Women from 25 to 45 years old who have dyed their hair blond in the last year.
- 18-50 year-olds who have bought rum in a supermarket of the brand of interest.
- Individuals between ages 25 and 65 who have cholesterol problems and consume cocoa powder.
- People who need orthodontic treatment but are not receiving it.
- Individuals who have drunk whiskey in the last three months, 25% of the brand of interest, 75% not against this brand.

Table 2 Population of interest: Proportion of surveys (of the 186 studied) with certain characteristics

Characteristic	%
Target population is the general population*	13.4
Target limited to only one gender	15.0
Target limited to some age group(s) (besides the 16+ or 18+)	52.1
More than one target of interest (within the same study)	19.3

* We count individuals between ages 16 (or 18) and 65 (or older) as the general population. For our purposes, the general Internet population is counted as the general population too.

acquire a sample of a population based on survey responses if respondents themselves do not know that they are part of the target population?

As we can see in Table 2, only 13.4% of surveys are interested in the general population, even very broadly defined (accepting age limits from 16 or 18 through 65 or older, and accepting the general Internet population). Furthermore, 15.0% of surveys are interested in only men or only women. 52.1% limit the population of interest to some age groups, besides the 16 or 18+ limit. 68.8% of surveys explicitly include a minimum age limit that spans from 8 to 55 years old. This limit is between 18 and 25 in 72.7% of cases. Besides, 47.8% of surveys include a maximum age limit that spans from 21 to 75 years old. The maximum age is 65 or older in 47.2% of cases. All this indicates that most surveys in the opt-in panel studied target

Table 3 The five most used quotas

Quotas on ...	Proportions of the surveys using these quotas (in%)
... Gender	78.5
... Age	72.6
... Geographical area	52.7
... Level of urbanization	8.7
... Social class	7.6

Note: For gender, N=158 because a quota is only possible when the population of interest includes both genders. For the others, N=186 (even for age, since even when the population is limited for some age groups, there are often still quotas within the rank of ages allowed)

very specific populations. In addition, 19.3% of the surveys define more than one target population, complicating matters even further.

Finally, most surveys also define some quotas. The goal of these quotas is usually to guarantee that the sample will be similar to the target population with respect to certain predefined variables. However, as we have just seen, target populations are often very specific. Most of the time, this means that we do not know the composition of the target population in terms of the main socio-demographics variables usually used as quotas. For instance, what is the gender or age distribution of the population of "people who need orthodontic treatment but are not receiving it"? In some cases, researchers have some ideas based on previous research. In others, quotas are used to make the sample similar to the Internet population or the whole panel, even if this does not correspond to the population of interest. It is therefore unclear the extent to which quotas are truly useful in improving the representativeness of the sample as it relates to the target population. Still, quotas were used in 95.2% of our 186 surveys. Table 3 shows the five most used quotas, with the proportions of surveys using each of them.

The most used quota is gender (78.5%), followed by age (72.6%) and geographical area (52.7%). Then come level of urbanization and social class, though their proportions are much smaller (8.7% and 7.6%, respectively). Variables such as having children, education or occupation are used in less than 3% of the surveys. We should note that these results may be strongly related to the country studied. For instance, in Latin American countries, the proportions of surveys that use quotas for social class in the Netquest panels is much higher, as habit differences across social classes are usually larger in Latin America.

4.1.2 What are the Characteristics of the Questionnaires?

After considering these surveys' targets, we researched the questionnaires' characteristics, in terms of topics, question formats and estimated length.

Survey topic. Table 4 presents the proportions of surveys dealing with various topics.

Up to 29.0% of surveys studied concern food or beverages. This is by far the most common topic. Surveys on society or politics come in second at 14.0%, whereas 11.8% of the surveys are about health, 8.6% are about insurance or banks, and 7.5% are about media, the Internet or new technologies. 71.0% of the surveys fit into one of these five categories. Some of the topics were more concrete and did not require prior knowledge (e.g., food) whereas others were more abstract and could have been affected by the respondents' level of knowledge on the topic (e.g., politics).

Question formats. Table 5 shows the proportions of surveys that made use of different question formats, from most to least common.

83.9% of the surveys include at least one multiple-response question, in which the respondents can/must select all the options they want/that apply. 76.3% of the surveys include at least one grid. Although some earlier research recommends avoiding multiple-response questions and grids (cf. Section 3.2), our study suggests that both are still very present. This is also true, in a lower proportion, for agree/disagree questions, which are present in 39.2% of the surveys. Three more formats are found quite frequently: open-response questions asking for a text answer (35.5% of the surveys), ordering questions (23.1%), and drop-down menus (18.3%). Again, this is the case despite evidence against these formats in academic literature. However, in the last case, further analyses would be needed to identify exactly which questions were asked using drop-down menus; in most cases it may only be a question on the province where the respondents live. Finally, videos are present in 7.5% of the surveys and sliders in only 2.7%. It is interesting to see that although web

Table 4 Main survey topics

Main topics	Proportions of the surveys within this topic (in%)
Food / Beverages	29.0
Society / Politics	14.0
Health	11.8
Insurance / Bank	8.6
Media / Internet / New Technologies	7.5
Others	29.1

Table 5 Proportions of surveys including different questions formats

Proportions of the surveys (in %) with at least one ...	
... Multiple-response question	83.9
... Grid	76.3
... Agree/Disagree question	39.2
... Open-text question	35.5
... Ordering question	23.1
... Drop-down menu	18.3
... Video	7.5
... Slider	2.7

surveys may allow these new features, they are not used much in practice in a panel like Netquest.

This overview shows that there is a clear gap between the academic guidelines on which question formats are best and which are actually used in online opt-in panel surveys. This suggests that the link between academic findings and their application in the practice of web survey administration must be improved. This overview also provides an indication of where further research could be useful, in order to study question formats that are often used in practice: for instance, even if forced-choice questions are recommended (Smyth et al. 2006; Revilla, 2015), more research about the evaluation of the quality of multiple-response questions could be useful, as this format continues to be used often.

Estimated survey length. In web surveys, completion time can vary greatly from respondent to respondent. Indeed, length depends on the rate at which respondents can read, process information, and answer questions; on the device used and the respondent's familiarity with the device, the speed of the Internet connection, the presence or frequency of interruptions, whether the respondent is multi-tasking or not, and so on. It also depends on the presence of filter questions. The estimated survey length (in minutes) can actually be very different from a given respondent's actual completion time. Thus, the estimated survey length gives more of an idea of the estimated complexity of the survey itself than of the experience of a given respondent. For this reason, we examine estimated survey length in this subsection on questionnaires' characteristics. Table 6 displays the minimum, maximum, average and median estimated survey length across all 186 surveys studied, as well as the proportions of surveys of different lengths.

Across all 186 surveys, the shortest had an estimated length of one minute, whereas the longest had an estimated length of 40 minutes. The average was 12

Table 6 Estimated length of the surveys in minutes

		Proportions (in %) of surveys with estimated length of ...	
Minimum across all surveys	1 minute	... 1-4 minutes	8.6
Maximum across all surveys	40 minutes	... 5-9 minutes	26.5
		... 10-14 minutes	30.3
Average for all surveys	12 minutes	... 15-19 minutes	16.8
		... 20-24 minutes	10.8
Median for all surveys	10 minutes	... 25-29 minutes	3.8
		... 30-40 minutes	3.2

minutes and the median was 10 minutes. This is much shorter than the average length across the surveys studied by Cape and Phillips (2015), which is 23 minutes in 2015. Cape and Phillips (2015) mention that the average adult attention span is around 20 minutes, and that 20 minutes is often considered the maximum questionnaire length for web surveys. This rule of thumb is actually used commonly.[2] In our study, 82.2% of surveys' have an estimated length of below 20 minutes, and only 7% have an estimated length of 25 minutes or more. Overall, the surveys in the panel we studied are quite short. We should mention, however, that the estimated length of the surveys programmed by Netquest (our focus here) is normally shorter than the estimated length of surveys sent to Netquest panelists but programmed directly by Netquest's clients. Thus, the average length would be slightly higher if we considered all surveys sent to Netquest panelists.

4.1.3 What did Respondents Get in Exchange for their Participation?

In general, the number of points respondents receive as an incentive corresponds to the estimated length in minutes, plus two. However, if the survey's estimated length is greater than 25 minutes, the incentive is increased further. Additional points are also sometimes awarded if a survey has specific requirements, such as two-wave surveys in which the researcher wants panelists to participate twice. Consequently, the correlation between estimated length and incentivization is very high, but

2 We found this rule discussed in many posts online, although some posts also discuss the pertinence of such a rule of thumb. See, for instance: http://blog.questionmark.com/how-many-questions-should-you-have-in-a-web-survey or http://researchaccess.com/2013/12/survey-length/.

Table 7 Incentives received (in number of points)

		Proportions (in %) of surveys with incentives of...	
Minimum across all surveys	4 points	... 4 points	0.5
Maximum across all surveys	58 points	... 5-9 points	25.3
		... 10-14 points	38.7
Average for all surveys	14 points	... 15-19 points	17.7
		... 20-24 points	9.7
Median for all surveys	12 points	... 25-29 points	5.4
		... 30-58 points	2.7

lower than one (around .95). Table 7 gives more information about the incentives received, in points.

Incentives span from four to 58 points, with an average of 14 and a median of 12. The highest proportion of surveys (38.7%) has an incentive between 10 and 14 points. To give some orientation on these points' value, we could mention that, for example, a panelist can acquire an e-book for 20 points, an online film for 40 points, a movie theater ticket for 120 points, or an eight gigabyte pen drive for 165 points.

4.1.4 What Happens During the Fieldwork? From Invitation to Completion

The participation process. Once the target population, questionnaire characteristics, and incentives are defined, panelists are invited to participate in the survey. In Netquest, profiling information is used to invite those panelists who are most likely to fit the target population. When these target populations are very specific, many panelists may need to be invited so that the study ultimately has enough respondents who fit the desired profile. In addition, most surveys use quotas: if the quotas are full, participants may be excluded. Finally, respondents might decide to abandon the survey, because they are experiencing some problems, because they do not like it, or simply because they have other things to do and forget to return to the survey. Table 8 provides further information on each step of the process, from invitation through completion. It also gives information about the participation, screen-out and break-off rates.

Table 8 shows major differences across surveys. The minimum number of invitations is 220, the maximum is 28,062, and the median is 2,239. Of the panelists invited, a minimum of 164 started, with a maximum of 18,019 and a median of

Table 8 Survey participation: from invitation to completion

	N	Minimum	Maximum	Average	Median
Number invited	182	220	28,062	3,437	2,239
Number started	182	164	18,019	2,131	1,450
Number screened out	185	1	14,291	1,105	466
Number of break-offs	185	2	2,261	131	49
Number of completes	186	90	5,015	875	602
Participation rate: (number of started / number of invited)*100	182	37.3	90.7	63.4	64.5
Screen-out rate: (number of screened out/number of started)*100	181	0.1	90.8	43.6	39.4
Break-off rate: [number of break-offs / (number of completes+ number of break-offs)]*100	185	1.1	62.1*	11.8*	6.7

Note: * These numbers are obtained excluding one special case: a survey where a product is sent to panelists' residences for testing. If we would include this survey, the maximum would be 88.9%, the average would be 12.2%, and the median 6.7%.

1,450. Then, 1 to 14,291 panelists were screened out, with a median of 466, and 2 to 2,261 break-offs, with a median of 49. In the end, 90 to 5,015 completed the full survey, with a median of 602.

The participation rate spans from 37.3% to 90.7%, with a median of 64.5%. Cross-survey differences are even more pronounced when considering the screen-out rate, which spans from 0.1% to 90.8%. This is related to the specificity of the target populations and to the profiling information available during sample selection. Overall, the median screen-out rate is high (39.4%) even if the company uses profiling information. This is an important problem for a fieldwork company because: a) it can affect the panelist's satisfaction and willingness to continue participating in the panel; indeed, it can be frustrating to discover you have been screened out; b) if panelists are rerouted to a profiling module, as they are in Netquest surveys, then the company must award them points even though they are screened out of the initial survey, resulting in a significant increase in costs; c) it is preferable not to send too many invitations to the same panelists; for instance, using opt-in panels data in the Nordic region, Stenbjerre and Laugesen (2005) found that six to 12 invitations per year is the frequency that leads to the best participation levels. However, if many panelists are screened out, more invitations must be sent to achieve a similar final number of completes. This can lead to panel overuse.

Finally, the break-off rate is low across the board, with a median of 6.7% and an average of 11.8% (excluding one survey where a product is sent to panelists' residences for testing). This is lower than the averages reported in many other studies: 30% for "general invitation surveys" and 15% for "individually targeted web surveys" (Galesic, 2006, p. 313). However, here again, there are huge variations across surveys (from 1.1% to 62.1%). We will try to explain these differences in Section 4.2 by examining the relationships between this break-off rate and various survey characteristics.

Number of sessions. Panelists can complete the questionnaire all at once or in several sessions, that is stopping and coming back later to continue. For respondents who completed the entire questionnaire ("completes"), the average number of sessions per survey is generally quite close to 1 (with a 1.2 average and a 1.1 median across all surveys), though it varies across surveys, from 1.0 to 2.9.

***Participation devices*.** For the "completes," we also had information on the type of device used to answer the survey. If respondents completed the survey in more than one session, we have information on only the device used in the first session. Table 9 gives the minimum, maximum, average and median proportions of PCs, tablets and smartphones used to participate across the 186 surveys.

Overall, PCs are still the main device of completion, but smartphone participation is not negligible. On average, across all surveys, 68.8% of panelists used a PC, 5.9% used a tablet and 25.2% used a smartphone. Again, there are large differences between surveys: some still have no mobile participation at all, whereas others have up to 39.7% tablet participation and 52.7% smartphone participation. It should be clear that surveys with no mobile participation at all are surveys in

Table 9 Proportion (in %) of different device types

Device	Minimum across all surveys	Maximum across all surveys	Average for all surveys	Median for all surveys
PC	40.9	100.0	68.8	67.4
Tablet	0	39.7	5.9	6.2
Smartphone	0	52.7	25.2	25.5

Table 10 Evaluation of the survey: average on a scale from (1) very badly done to (5) very well done

	Minimum across all surveys	Maximum across all surveys	Average for all surveys	Median for all surveys
All completes	3.2	4.5	4.0	4.1
PC respondents only	3.3	4.6	4.1	4.1
Smartphone respondents only*	3.1	4.5	4.0	4.0

* N=170 in the case of smartphone respondents only, as the surveys with no or very few smartphone respondents are not considered in this row.

which mobile devices were not allowed; if mobile devices were always allowed, their usage rate would be higher.

4.1.5 Evaluation of the Survey

The last step in this overview is to examine how respondents evaluated the surveys they took. Respondents who completed the entire survey were prompted to answer a final question: "Finally, what did you think about this survey? Select from 1 to 5 stars to indicate if you think the survey was (1) very badly done to (5) very well done." In this case, we also had information that allowed us to look at the evaluation for PC respondents only and for smartphone respondents only. We did not consider tablets only, because of the low rate of tablet usage. Table 10 reports the results across the 186 surveys.

The survey evaluations overall are quite positive: the median across the 186 surveys is 4.1 on a scale from 1 to 5. Moreover, variation across surveys is small, with a minimum of 3.2 and a maximum of 4.5. The results are similar when considering only PC and only smartphone respondents.

4.2 How Break-off Levels and Survey Evaluations Relate to Key Questionnaire Characteristics

In this section, we study how respondents' break-off levels and survey evaluations relate to key questionnaire characteristics, namely 1) their topic (dummies for the five main topics, "other" being the reference category), 2) the presence of different question formats (eight dummies; 1 meaning that the format is present at least once) and 3) the estimated survey length (continuous variable).

Looking at the distribution of the break-off rate across the 186 surveys shows a very skewed distribution toward the left, with two outliers on the right side (see Appendix 1a). In addition, the residuals of a simple OLS regression are not normally distributed (Appendix 1b). In order to deal with this, we use the logarithm of the break-off rate as a dependent variable.[3] This helps resolve the problem of the outliers, non-normality of the residuals (Appendix 1c), and heteroscedasticity (p=.49 for the Breusch-Pagan test with the transformed variable).

Concerning the survey evaluation, we use the average evaluation across all respondents in a given survey as a dependent variable; this takes 14 different values, ranging from 3.2 to 4.5. In this case, examination of the standardized normal probability plot suggests that residuals approximate normal distributions (Appendix 1d), so we use an OLS regression.[4] The results are presented in Table 11.

First, in the case of break-off, the whole model is significant (p<.01), and the model explains 28.2% of the variance (adjusted R^2). However, only two variables have significant effects (5% threshold): the presence of at least one video and the estimated questionnaire length.

In the survey evaluation model, the whole model is not significant (p=.15) and the explained variance is very low (Adjusted R^2 =.0308). None of the variables has a significant effect, suggesting that the survey characteristics affect the break-off rate more than the survey evaluation conducted by those panelists who completed the survey.

3 Other approaches have been tested, namely: a) a Poisson regression with robust variance and excluding the two outliers (two surveys with much higher break-off rates than all others): this led to similar conclusions; b) a negative binomial regression also excluding the two outliers: in this case, besides the estimated length and presence of at least one video (which are significant in the results presented here), the presence of at least one slider and at least one grid also had significant impacts on the break-off.

4 We also ran an ordered logistic regression, and a regression with exponential or log transformation: the conclusions remain the same.

Table 11 Regressions of log(break-off rate) and of survey evaluation on survey characteristics

		Log(break-off rate) *N=184*			Survey evaluation *N=185*		
Explanatory variables		Coef.	S.E	P-value	Coef.	S.E	P-value
Survey Main Topic	Food / Beverages	.21	.17	.21	.05	.05	.26
	Society / Politics	-.06	.21	.76	-.04	.06	.46
	Health	-.02	.22	.92	-.01	.06	.93
	Insurance / Bank	-.18	.25	.46	<.01	.07	.98
	Media / Internet / New technologies	-.11	.26	-.42	-.03	.07	.69
Format of questions	Includes one or more slider	.63	.39	.11	.09	.11	.45
	Includes one or more ordering question	-.01	.15	.95	.01	.04	.83
	Includes one or more grid	.21	.16	.20	.02	.05	.69
	Includes one or more agree/ disagree question	.19	.14	.18	-.06	.04	.16
	Includes one or more multiple-response question	-.15	.19	.44	.09	.05	.09
	Includes one or more video	.65	.25	.01	.02	.07	.83
	Includes one or more open-text question	.18	.14	.19	-.06	.04	.12
	Includes one or more drop-down menu	.03	.17	.85	-.04	.05	.43
Survey length	Estimated length	.06	.01	<.01	<.01	<.01	.68
	Constant	1.05	.24	<.01	4.02	.07	<.01
Model fit	R^2		.3374			.1045	
	Adj. R^2		.2825			.0308	

5 Discussion and Conclusions

In this study, we have reviewed all of the surveys programmed by the online fieldwork company Netquest and implemented in their panel in Spain over a period of approximately six months. By reviewing several aspects of the 186 surveys, we are able to highlight the true practice of web surveys in a current opt-in online panel. Some of the main results that we wish to emphasize are presented below.

- Survey target populations: these are often very specific, and previous information about them is hard to come by, making it difficult to use quotas that ensure the sample is similar to the population of interest.
- Survey topic: 29% of the surveys are about food or drinks. This is the most common topic. The five most common topics cover 71.0% of all surveys.
- Question format: multiple-response questions are used very frequently, as are grids. Agree/disagree questions, open-text questions, ordering questions, and drop-down menus are also used quite often. On the other hand, videos and sliders are present in less than 10% of surveys.
- Estimated length: 56.8% of surveys have an estimated length between 5 and 14 minutes.
- Incentives: on average, the incentive for answering one survey is 14 points.
- Participation process: given the specificity of target populations and the use of quotas, it is sometimes necessary to invite a huge number of panelists in order to attain a small final number of completes. However, screen-out rates vary widely across surveys. The break-off rate also varies widely, but is much smaller overall. We are able to identify some variables that seem to be related with a higher break-off, including the presence of one or more videos and a longer estimated length. Thus, we would recommend avoiding videos and keeping questionnaires as short as possible.
- Devices used: although PCs are the main device used for participation, mobile participation is clearly non-negligible.
- Survey evaluation: the survey evaluation does not vary much across surveys, and is also very similar for PC and smartphone respondents. As variations are minimal, it is not surprising that we did not find any significant effect in the regression analysis.

Our overview suggests that opt-in panels are very distinct from other web surveys, in terms of the population they attempt to cover. This also has an effect on the participation process, and in particular on the screen-out rate. Opt-in panels can differ in other respects, too. Further research focusing on these opt-in panels is needed in order to better understand the specific challenges that they face, and the best approaches to overcome those challenges. This study was limited to only one panel, in one country, and we could only analyze variables at the survey level. In

order to further study the reasons for break-off, information about what takes place just before the break-off, rather than characteristics of the survey as a whole, is needed. This study was not able to account for the device used to complete the survey in many of the analyses presented, even if it would have been very interesting, for example, to study the break-off separately for PC and smartphone respondents. Future research in these directions would be helpful.

However, even more than a need for further research, this study suggests that there is a gap between research and practice, particularly in relation to question formats. Indeed, the guidelines from academic research recommend that question formats like multiple-response questions and grids should be avoided, but our analyses in this overview reveal that they are still used very often. Academic researchers may need to work harder when sharing their results and convincing practitioners to follow their recommendations. Researchers may also need to further adapt their research so it better meets practitioners' needs. To achieve these goals, they may need to look more closely at the reality of conducting surveys in the 21st century and focus further research on how to improve the most frequently used question formats.

References

AAPOR Standards Committee (2010). *AAPOR report on online panels*. AAPOR Executive Council. doi:10.1093/poq/nfq048

Bethlehem, J. & Biffignandi, S. (2011). *Handbook of Web Surveys*. NY: Wiley. ISBN: 978-0-470-60356-7

Brace, I. (2004). *Questionnaire design: how to plan, structure and write survey material for effective market research*. London 2004.

Callegaro, M. (2010). Do You Know Which Device Your Respondent Has Used to Take Your Online Survey? *Survey Practice*.

Callegaro, M., Baker, R.P., Bethlehem, J., Göritz, A.S., Krosnick, J.A., & Lavrakas, P.J. (2014). *Online panel research. A data quality perspective*. Chichester: Wiley.

Callegaro, M., Lozar Manfreda, K., & Vehovar, V. (2015). *Web Survey Methodology*. London: Sage.

Cape, P. (2012). Understanding Respondent Motivation. DGOF White paper, available at: http://www.dgof.de/wp3/wp-content/uploads/2012/07/WP_Understanding-Respondent-Motivation.pdf

Cape, P. & Phillips, K. (2015). Questionnaire Length and Fatigue Effects: The Latest Thinking and Practical Solutions. White paper, available at: https://www.surveysampling.com/site/assets/files/1586/questionnaire-length-and-fatiigue-effects-the-latest-thinking-and-practical-solutions.pdf

Chan, P., & Ambrose, D. (2011). Canadian online panels: Similar or different? *Vue Magazine*, 16-20. Retrieved from: http://www.mktginc.com/pdf/VUE%20JanFeb%202011021.pdf.

Couper, M.P. (2000). Web surveys – A review of issues and approaches. *Public Opinion Quarterly*, 64, 464-494.

Couper, M.P. (2008). *Designing effective web surveys*. NY: Cambridge University Press.

Couper, M.P, & Miller, P.V. (2008). Introduction to the special issue. *Public Opinion Quarterly*, 72(5), 831-835 doi:10.1093/poq/nfn066

Couper, M.P., Traugott, M., & Lamias, M. (2001). Web survey design and administration. *Public Opinion Quarterly*, 65(2), 230-253. doi:10.1086/322199

De Bruijne, M., & Wijnant, A. (2014). Improving Response Rates and Questionnaire Design for Mobile Web Surveys. *Public Opinion Quarterly*, 78(4): 951-962.

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

Dillman, D.A. (2000). *Mail and internet surveys: The tailored design method* (2nd ed.). New York: John Wiley & Sons.

Dillman, D.A., & Bowker, D.K. (2001). The Web questionnaire challenge to survey methodologists. In U.-D. Reips, & M. Bosnjak (Eds.), *Dimensions of Internet science* (p. 159-178). Lengerich: Pabst Science Publishers.

Dillman, D.A., Smyth, J.D., & Christian, L.M. (2009). *Internet, Mail, and Mixed-Mode Surveys: The Tailored Design Method*. New York: Wiley.

Fowler, F.J. (1995). Improving Survey Questions: Design and Evaluation. *Applied Social Research Methods Series*, 38, 56-57.

Funke, F., Reips, U.-D., & Thomas, R.K. (2011). Sliders for the Smart: Type of Rating Scale on the Web Interacts with Educational Level. *Social Science Computer Review*, *29*(2), 221-231. doi:10.1177/0894439310376896

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

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

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

Healey, B. (2007). Drop Downs and Scroll Mice: The Effect of Response Option Format and Input Mechanism Employed on Data Quality in Web Surveys. *Social Science Computer Review*, 25(1), 111-128. doi: 10.1177/0894439306293888

Herzog, A.R & Bachman, J.G. (1981). Effects of questionnaire length on response quality. *Public Opinion Quarterly*, 45, 549-559.

Krosnick, J.A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology*, 5, 213-236.

Lambert, A.D., & Miller, A.L. (2015). Living with Smartphones: Does Completion Device Affect Survey Responses? *Research in Higher Education,* 56, 166-177

Lorch, J., & Mitchell, N. (2014). Why You Need to Make Your Surveys Mobile Friendly Now. Webinar organized by the American Marketing Association, May 2014, https://www.ama.org/multimedia/Webcasts/Pages/why-you-need-to-make-your-surveys-mobile-friendly-now-051514.aspx?tab=home

Lozar Manfreda, K.L, Batagelj, Z., & Vehovar, V. (2002). Design of Web survey questionnaires: Three basic experiments. *Journal of Computer-Mediated Communication*, 7(3), 0, doi:10.1111/j.1083-6101.2002.tb00149.x

Marcus, B., Bosnjak, M., Lindner, S., Pilischenko, S., & Schütz, A. (2007). Compensating for Low Topic Interest and Long Surveys: A Field Experiment on Nonresponse in Web Surveys. *Social Science Computer Review,* 25(3), 372-383.

Matthijsse, S.M., de Leeuw, E.D., & Hox, J.J. (2015). Internet panels, professional respondents, and data quality. *Methodology*, 11(3), 81-88 Article doi:10.1027/1614-2241/a000094

Oppenheim, A.N. (1992). *Questionnaire design, interviewing and attitude measurement*. London: Pinter Publishers.

Ovadia, S. (2004). Ratings and rankings: Reconsidering the structure of values and their measurement. *International Journal of Social Research Methodology*, 7, 403-14.

Parsons, C. (2007). Web-Based Surveys: Best Practices Based on the Research Literature. Visitor Studies, 10(1), 13-33, doi:10.1080/10645570701263404.

Peytchev, A., & Hill, C.A. (2010). Experiments in mobile web survey design: Similarities to other modes and unique considerations. *Social Science Computer Review*, 28, 19-335

Poynter, R. (2001). A Guide to Best Practice in Online Quantitative Research. In A. Westlake, W. Sykes, T. Manners, & M. Rigg (Eds.), *The Challenge of the Internet* (pp. 3-19); Proceedings of the ASC International Conference on Survey Research Methods. London: Association for Survey Computing.

Revilla, M. (2015). Effect of using different labels for the scales in a web survey. *International Journal of Market Research,* 57(2), 225-238. doi:10.2501/IJMR-2014-028

Revilla, M., Toninelli, D., Ochoa, C., & Loewe, G. (2016). Do online access panels really need to allow and adapt surveys to mobile devices? *Internet Research*, 26(5), 1209-1227

Saris, W. E., & Gallhofer, I. (2014). *Design, evaluation, and analysis of questionnaires for survey research*. New York: JohnWiley & Sons, Inc. (Second Edition).

Saris, W.E., Revilla, M., Krosnick, J.A., & Shaeffer, E.M. (2010). Comparing Questions with Agree/Disagree Response Options to Questions with Construct-Specific Response Options. *Survey Research Methods*, 4(1), 61-79.

Schuman, H., & Presser, S. (1981). *Questions and answers in attitude surveys: Experiments in question form, wording, and context*. New York: Academic Press.

Smyth, J.D., Dillman, D.A., Christian, L.M., & Stern, M.J. (2006). Comparing check-all and forced-choice question formats in web surveys. *Public Opinion Quarterly,* 70(1): 66–77.

Stenbjerre, M., & Laugesen, J.N. (2005). Conducting Representative Online Research: A Summary of Five Years of Learnings. Paper presented at ESOMAR Worldwide Panel Research Conference in Budapest, April 17-19 2005. Available at: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.485.1158&rep=rep1&type=pdf

Sudman, S., & Bradburn, N.M. (1982). *Asking questions*. San Francisco: Jossey-Bass Inc Pub.

Tourangeau, R., Conrad, F.G., & Couper, M.P. (2013). *The Science of Web Surveys*. Oxford University Press. ISBN: 978-0-19-974704-7

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

Vonk, T., Van Ossenbruggen, R., & Willems, P. (2006). The effects of panel recruitment and management on research results: A study across 19 panels. ESOMAR: *Panel Research 2006*, 79-100

Walker, R., Pettit, R., & Rubinson, J. (2009). The foundations of quality initiative: a five-part

immersion into the quality of online research. Journal of Advertising Research, 49(4), 464-485.

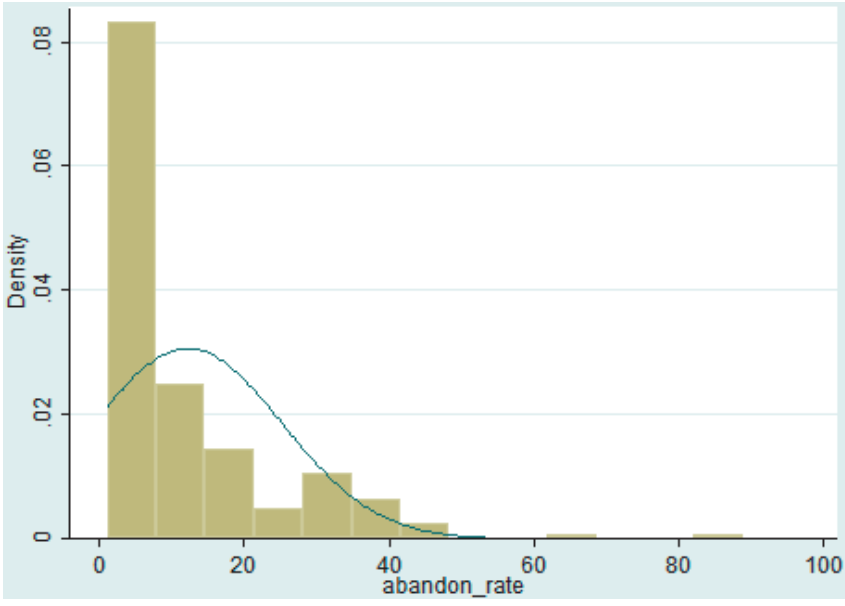
Wells, T., Bailey, J.T., & Link, M.W. (2013). Filling the void: Gaining a better understanding of tablet-based surveys. *Survey Practice*, 6

Yan, T., Conrad, F.G., Tourangeau, R., & Couper, M.P. (2010). Should I Stay or Should I Go: The Effects of Progress Feedback, Promised Task Duration, and Length of Questionnaire on Completing Web Surveys. *International Journal of Public Opinion Research*, 23(2), 131-147. doi:10.1093/ijpor/edq046.

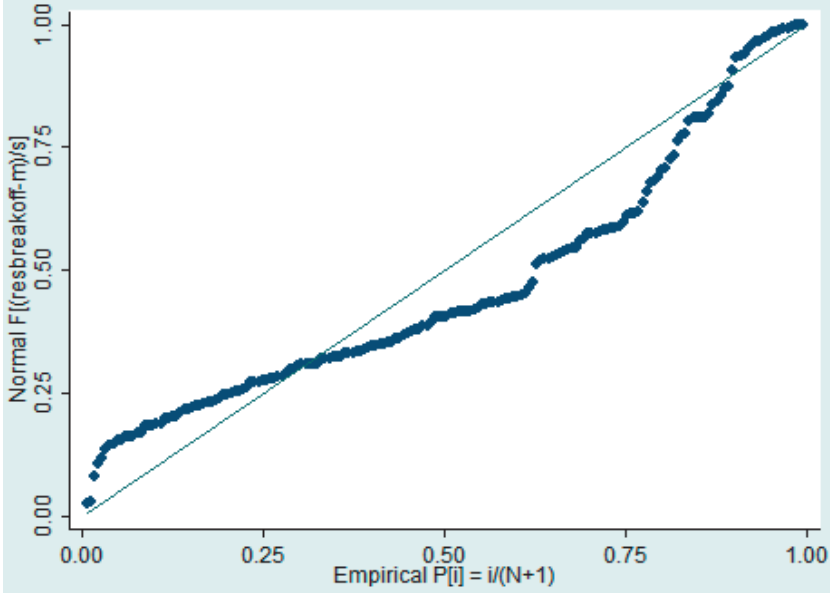
Zhang, C., & Conrad, F. (2014). Speeding in Web Surveys: The Tendency to Answer Very Fast and Its Association with Straightlining. *Survey Research Methods, 8*(2), 127-135.

Appendix 1

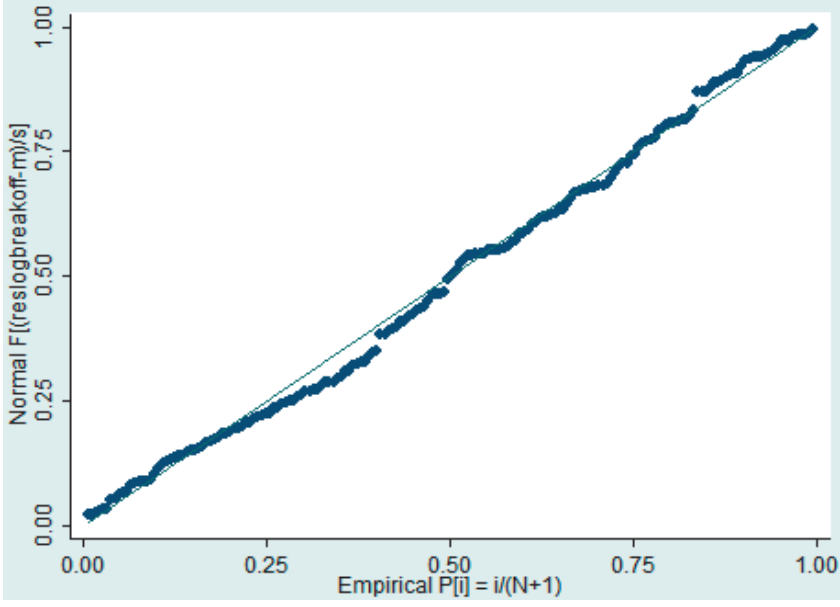
a) Histogram of the break-off rate



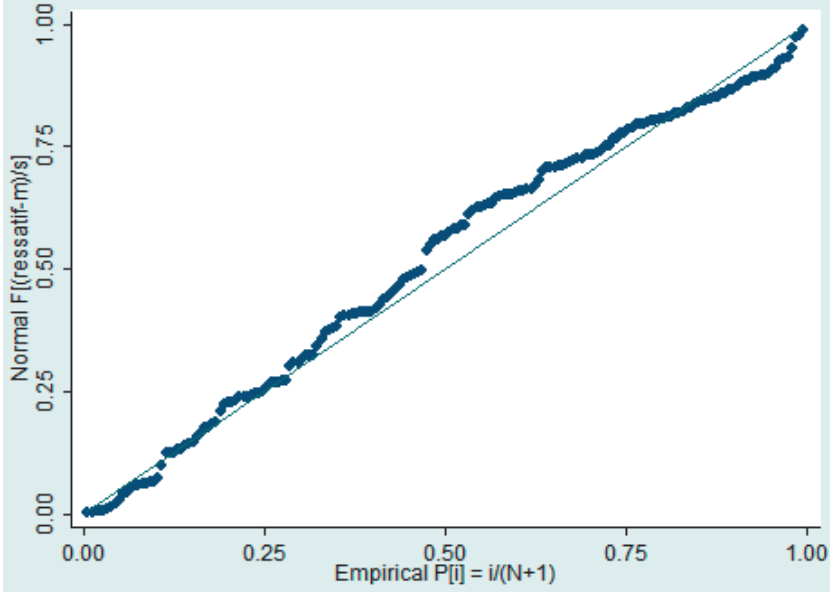
b) P-P plot of the residuals when using break-off rate



c) P-P plot of the residuals when using the logarithm of break-off rate



d) P-P plot of the residuals for average survey evaluation



methods, data, analyses | Vol. 11(2), 2017, pp. 163-188 DOI: 10.12758/mda.2017.03

How Interviewer Effects Differ in Real and Falsified Survey Data: Using Multilevel Analysis to Identify Interviewer Falsifications

Uta Landrock
University of Kaiserslautern

Abstract

In face-to-face interviews, interviewers can have an important positive influence on the quality of survey data, but they can also introduce interviewer effects. What is even more problematic is that interviewers may decide to falsify all or parts of interviews. The question that the present article seeks to answer is whether the interviewer effects found in falsified data are similar to those found in real data, or whether interviewer effects are larger and more diverse in falsified data and may thus be used as an indicator for data contamination by interviewer falsifications. To investigate this question, experimental data were used from controlled real interviews, interviews falsified by the same interviewers, and questionnaires completed by these interviewers themselves as respondents. Intraclass correlations and multilevel regression models were applied, and interviewer effects in the real survey data were compared with those in the falsified data. No evidence of interviewer effects was found in the real data. By contrast, interviewer effects were found in the falsified data. In particular, there was a significant association between the interviewers' own responses and the falsified responses to the same questions in the questionnaire. Thus, to detect interviewer falsifications, I recommend that researchers should also get the interviewers to complete the questionnaire and check datasets or suspicious cases for interviewer effects.

Keywords: interviewer, interviewer effects, interviewer falsifications, data quality, identification of falsifications, multilevel analysis

© The Author(s) 2017. This is an Open Access article distributed under the terms of the Creative Commons Attribution 3.0 License. Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.

1 Introduction

Face-to-face interviews are an important mode of data collection in empirical social research. It is used in many major studies, for example the European Values Study (EVS),[1] the U.S. General Social Survey (GSS),[2] and the Programme for the International Assessment of Adult Competencies (PIAAC).[3] Interviewers can have a major influence on the quality of survey data. On the one hand, they can improve data quality, for example by helping the respondent to understand the survey questions correctly (Mangione, Fowler, & Louis, 1992). On the other hand, there is the risk of interviewer effects, that is, distortions of survey responses due to the presence of an interviewer. Interviewer effects can cause biased data and affect substantive findings (Beullens & Loosveldt, 2016; Groves & Magilavy, 1986). They occur when the respondent's answer depends not only on the intended stimulus of the question but also on the interview situation and the interviewer (Bogner & Landrock, 2016; Schanz, 1981). In the case of interviewer effects, certain interviewer behaviors (e.g., reading pace or suggestiveness) or characteristics (e.g., experience, age, gender, or education) may influence the response behavior of the respondent (Beullens & Loosveldt, 2016; Haunberger, 2006; Mangione et al., 1992). Interviewer effects therefore constitute response bias (see Groves & Magilavy, 1986), where the reported values of the respondent systematically deviate from the true values.

In this context, it is important to know whether some types of questions are more susceptible to interviewer effects than others (Mangione et al., 1992). Research on interviewer effects has yielded a large number of findings in this regard (for an overview, see Bogner & Landrock, 2016). According to Haunberger (2006), for example, difficult and sensitive questions, attitudinal questions, and open-ended questions are particularly prone to interviewer effects. Haunberger (2006) showed that, in the case of difficult questions, the gender and education of the interviewers may have an influence on responses, for example, to income-related questions. The probability that the respondent will refuse to answer such questions is reported to

1 http://www.europeanvaluesstudy.eu/

2 http://gss.norc.org/

3 http://www.oecd.org/skills/piaac/

Acknowledgements

The research presented in this paper was funded by a German Research Foundation (DFG) grant awarded for the project IFiS – Identification of Falsifications in Surveys (WI 2024/5-4 and ME 3538/4-1). This financial support is gratefully acknowledged. I would also like to thank the referees and the editors for their helpful comments on an earlier version of the manuscript.

Direct correspondence to

Uta Landrock, University of Kaiserslautern, Erwin-Schroedinger-Straße 57, 67653 Kaiserslautern, Germany
E-mail: landrock@sowi.uni-kl.de

be higher in the case of female or highly educated interviewers (Bogner & Landrock, 2016; Haunberger, 2006). Regarding attitudinal questions, research findings are ambiguous. Whereas Liu and Stainback (2013) identified interviewer gender effects on responses to attitudinal questions, Groves and Magilavy (1986) did not find evidence of such an influence on attitudinal questions compared to factual questions. Haunberger (2006) suggested that interviewer age and education may influence responses to open-ended questions and that these questions are therefore susceptible to interviewer effects (Mangione et al., 1992). By contrast, Groves and Magilavy (1986) reported that open-ended questions were not inherently more susceptible to interviewer effects than closed questions. However, in the case of open questions that ask respondents to mention several entities, for example "What do you think are the most important problems facing the country?," the authors suggested that the likelihood that the respondent would mention a second entity might depend on the interviewer's probing behavior, and that "the differential behaviors that determine whether a second mention is given also might influence substantive responses on the second mention" (Groves & Magilavy, 1986, p. 260). In summary, therefore, research findings show that difficult, attitudinal, and open-ended questions are susceptible to interviewer effects.

These findings provide evidence that the perceptible sociodemographic characteristics of the interviewer – namely gender, age, and education – are relevant to the occurrence of interviewer effects (Haunberger, 2006; Liu & Stainback, 2013; West & Blom, 2016). Olson and Bilgen (2011) reported that larger interviewer effects occurred with respect to acquiescence in the case of experienced interviewers than in the case of inexperienced interviewers. West and Blom (2016) described the influence of certain personality traits of the interviewers that may affect response behavior. Moreover, research findings suggest that the relation between interviewers' and respondents' characteristics may result in interviewer effects: Schanz (1981) analyzed the relevance of interaction effects and described positive correlations between the answers of the interviewer and the answers of the respondent to the same survey questions. One possible explanation for this positive correlation is that the respondent reacts to the non-verbally expressed attitudes of the interviewer (Schanz, 1981; West & Blom, 2016). Thus, interviewer effects may also depend on the content of the question and the interaction of the attitudes of the interviewers and the respondents (Schanz, 1981).

In face-to-face interviews, not only may interviewer effects occur, but interviewers may even decide to falsify all or parts of interviews. This is the most extreme and problematic form of influence that an interviewer can exert. Falsifications may severely bias the results of analyses and lead to incorrect results (Landrock, 2017; Reuband, 1990; Schnell, 1991; Schraepler & Wagner, 2003). A reliable strategy for identifying falsifications would therefore be extremely valuable to ensure high quality in interviewer-based survey research. However, research has

shown that, based on univariate distributions (Menold & Kemper, 2014; Reuband, 1990; Schnell, 1991) and multivariate correlations (Landrock, 2017), falsified and real data appear to be quite similar and that the existence of falsifications in data is thus not readily noticeable. Given that the falsification of interviews may be considered to be an extreme form of interviewer effect, statistically testing for interviewer effects might provide a more effective indicator for identifying falsifications. This paper therefore analyzes and compares interviewer effects in real survey data and in data falsified by interviewers. Using experimental data, the aim is to determine whether similar interviewer effects occur in falsified data and in real data or whether interviewer effects are larger and more diverse in falsified data and may thus be used as an indicator for data contamination by interviewer falsifications (see Winker, Kruse, Menold, & Landrock, 2015).

In falsified interviews, by definition, no interaction takes place between the respondent and the interviewer. Therefore, it may seem implausible to assume that interviewer effects occur in a dataset comprised of falsified data. However, in falsified interviews, interviewers obviously have a direct influence on the data reported as answers by the respondent. Yet, they have only a little information about the respondent. Consequently, the fabrication of plausible responses depends very strongly on the falsifier. Thus, interviewer effects – or, more precisely, "falsifier effects" – can be expected.

Different falsifiers may falsify the respondents' answers in different ways. It is conceivable that certain socioeconomic, demographic, or psychological characteristics of the falsifiers may find their way into the data they falsify. Both the falsifiers' perceptions of social reality and their falsifications are influenced by personal characteristics. Therefore, the interviewers' characteristics should be significant explanatory variables in a dataset that is contaminated by interviewer falsifications. Moreover, I assume that interviewer effects are more pronounced in falsified than in real survey data (see Winker et al., 2015).

In the research presented in this paper, a number of variables that are known to be generally susceptible to interviewer effects are analyzed as dependent variables with the aim of determining (a) the degree to which interviewer effects occur in real and in falsified data and (b) whether there are differences between the interviewer effects in real and in falsified survey data.

2 Hypotheses

To contribute to research on interviewer effects, to knowledge of interviewer falsifications and their impact on data quality, and to potential strategies for identifying contaminated data, the following two general hypotheses will be tested:

H1: Interviewer effects occur both in real and in falsified data.

As falsifying interviewers have only a little information about the respondent, they must draw on their personal experience of social reality in order to fabricate plausible answers to survey questions. Thus, interviewer effects may occur not only in real survey data but also in falsified survey data (see Winker et al., 2015).

H2: The interviewer effects in falsified data are larger than in real data.

I assume that sociodemographic or psychological characteristics of interviewers are more likely to find their way into falsified survey data than into real data.

Regarding the interviewer characteristics that may cause interviewer effects or influence the way in which an interviewer falsifies, explanatory variables will be analyzed that can theoretically be expected to be susceptible to interviewer effects. The following more specific hypotheses will be tested on real data and on falsified data:

H3a: The core sociodemographic characteristics of the interviewers affect the reported responses.

As reported by West and Blom (2016), Haunberger (2006), Mangione et al. (1992), and Liu and Stainback (2013), sociodemographic characteristics of the interviewer – in particular gender, age, and education – may lead to interviewer effects. I further expect that income, as an indicator of socioeconomic background, may also cause interviewer effects.

H3b: The magnitude of interviewer effects depends on the interviewer's experience.

Olson and Bilgen (2011) found that experienced interviewers caused larger interviewer effects than inexperienced interviewers. Hypothesis H3b will test whether this finding is replicated in the present study.

H3c: Associations exist between the behaviors and attitudes of interviewers and the reported behaviors and attitudes of the respondents they interview.

Following Schanz (1981), I assume that associations will be found between the answers of the interviewers and the answers of the respondents to the same survey question – in other words, that the interviewer's response to the same survey question affects the response reported by the respondent.

H3d: The occurrence and magnitude of interviewer effects depends on the personality traits of the interviewer.

Both West and Blom (2016) and Winker et al. (2015) found evidence that suggested that the personality traits of the interviewer may lead to interviewer effects. West and Blom (2016) reported an effect of interviewers' extraversion and self-confidence. Accordingly, I assume that interviewers with higher levels of extraversion produce larger interviewer effects than introverted interviewers. By contrast, more

conscientious interviewers should produce smaller interviewer effects than interviewers with a lower level of conscientiousness. With regard to self-confidence, I assume that interviewers with a higher level of perceived self-efficacy perform better, and therefore produce smaller interviewer effects, than interviewers with a lower level of perceived self-efficacy.

H3e: The magnitude of interviewer effects depends on the interviewer payment scheme used (payment per completed interview vs. payment per hour).

In their study of interviewer effects in real and falsified interviews, Winker et al. (2015) found that the payment scheme (i.e., the type of monetary compensation) applied had an impact on the collected data and therefore on the quality of a survey. I assume that interviewers who are paid per completed interview produce larger interviewer effects than interviewers paid per hour. Winker et al. (2015) also found correlations between the payment scheme and political participation (operationalized as the number of political activities mentioned by the respondent). For the real data, the authors showed that payment per hour was associated with a higher number of political activities mentioned. It would appear that payment per hour leads to more complete data and thus to higher data quality. Hypothesis H3e will test the assumption that interviewers who are paid per completed interview produce larger interviewer effects than interviewers who are paid per hour.

3 Data Base and Methods

Due to the virtual non-existence of datasets with proven falsified interviews, experimental data were used to analyze falsified data and their differences to real data (see Winker et al., 2015). My data base comprised three datasets. The data were collected at the University of Giessen, Germany in summer 2011 in the framework of the research project IFiS – Identification of Falsifications in Surveys (see also Menold & Kemper, 2014; Winker et al., 2015).

In the first step, 78 interviewers conducted 710 real face-to-face interviews. The questionnaire consisted of 62 questions, which were taken mainly from the 1998 German General Social Survey (ALLBUS) questionnaire.[4] Besides sociodemographic questions, the questionnaire comprised attitudinal and behavioral items on social, political, and economic topics. The average interview duration was 30 minutes. Both the respondents and the interviewers were students at the University of Giessen. The interviewers themselves selected the respondents on the university campus without any quota restrictions and interviewed them. The audio-recorded interviews were checked to make sure that they had been conducted correctly. Half

4 http://www.gesis.org/en/allbus/allbus-home/

of the interviewers were paid per completed interview (8 euros), the other half were paid per hour (12 euros). Prior to data collection, an interviewer training session was conducted, in the course of which the interviewers were familiarized with the research design and the questionnaire.

For the second dataset, 710 interviews were fabricated. For this purpose, the same interviewers who had conducted the real interviews were requested to fabricate survey data in the lab. Hence, for each real interview, a corresponding fabricated interview was obtained. Compensation was allocated either per interview (3 euros per falsified interview) or per hour (9 euros per hour). The falsifying interviewers were given details of the sociodemographic characteristics of the persons whose interviews they were to fabricate. These persons were real survey participants, who had been interviewed previously by another student interviewer. The information provided included the respondent's gender, age, subject studied, number of semesters enrolled, marital status, place of residence, living situation (i.e., the person or persons with whom the respondent lived in a household), and country of origin. In the case of a genuine (i.e., uninstructed) falsification in an actual fieldwork setting, the falsifying interviewer could easily have obtained this information by briefly interviewing the respondent. The falsifiers were requested to imagine the described person and to complete the questionnaire, thus fabricating the data as if they had been collected in a real survey fieldwork setting.

The exact instructions for falsifying an interview were:

> Please read carefully the description of the person whose interview you are to falsify. Please complete the attached questionnaire as if you had really conducted a personal interview with the respondent. During falsification, please place the description of the respondent next to the questionnaire, so that you are always aware of the characteristics of that person.
>
> The person whose interview you are to falsify…
>
> - is female,
> - is 20 years old,
> - studies teaching,
> - is enrolled in her second semester at a university.
> - She is unmarried, in a steady relationship,
> - lives in Huettenberg, a rural village in Hesse,
> - with her parents or relatives.
> - Country of birth: Germany.

As a last step, the interviewers themselves, as respondents, completed the same questionnaire that they had previously used for interviewing and falsifying. These self-administered interviews generated the third dataset.

This experimental setup has strengths, but it also has weaknesses. One weakness is that the respondents and interviewers were students and that core sociodemographic characteristics, such as age and education, therefore displayed only small variance (see Winker et al., 2015). The major strength of the experimental setup, compared to a standard field setting, was the possibility of collecting more information about the interviewers and their falsifying processes. Because they were surveyed with the same questionnaire as the proper respondents, the dataset includes not only information about respondents and fictitious respondents but also about the interviewers. This offers great potential for analyzing interviewer effects.

There are several possible approaches to investigating interviewer effects. Schanz (1981) analyzed the influences of interviewer characteristics on the response behavior of the participants by estimating multiple regression analyses. First, he included substantive explanatory variables; then he added interviewer variables. Mangione et al. (1992) and Groves and Magilavy (1986) measured interviewer effects by intraclass correlation. The intraclass correlation expresses the proportion of the item variance that is attributable to the interviewer (Mangione et al., 1992). In the absence of interviewer effects, the value of the intraclass correlation should be zero or close to zero (Beullens & Loosveldt, 2016). Olson and Bilgen (2011) estimated multilevel regression analyses with respondent characteristics such as age and education on the respondent level (individual level) and interviewer characteristics such as age, education, and experience on the interviewer level (contextual level).

At first glance, it would appear to be useful to estimate ordinary least squares (OLS) regressions. However, especially when it comes to analyzing interviewer effects, it makes sense to assume that – as expressed in the above-mentioned hypotheses – the observations of the respondents (i.e., the individual interviews) are probably not independent from the interviewers. Therefore, the model assumptions of OLS regressions are not met. Rather, the data are organized hierarchically, and multilevel regression analyses are thus more appropriate (Hox, 1995). The respondents represent the individual level, and the interviewers represent the group or contextual level.

To investigate the impact of interviewer characteristics on substantive findings, intraclass correlations were also estimated and multilevel regression analyses were conducted. To answer the research question as to what influence interviewers have on the data and findings and whether there are differences between real and falsified data in this respect, identical multilevel regression models were estimated separately with real and with falsified data. Thus, to determine what differences occur, the respective results – in particular, the effects of the various independent variables – were compared. This approach also allowed the identification of interviewer effects on substantive findings.

Table 1 Overview of variables used to analyze interviewer effects

Dependent Variables	Independent Variables on the Individual (Respondent) Level	Independent Variables on the Contextual (Interviewer) Level
Income	Age Living situation	Payment scheme
Political participation	Gender Internal political efficacy Political dissatisfaction Extremism	Interviewer's gender Interviewer's income Interviewer's response to the same questions of the questionnaire
Political anomy	Economic dissatisfaction External political efficacy	Interviewer's experience
Healthy eating behavior	Intention Perceived behavioral control TV consumption Body mass index Doing sports Preference for healthy desserts	Interviewer's extraversion Interviewer's conscientiousness Interviewer's level of perceived self-efficacy

4 Operationalization and Multilevel Regression Model

Table 1 gives an overview of the dependent and independent variables used. These variables are explained in more detail in the following sections.

4.1 Dependent Variables on the Individual Level

One aim of the present study was to analyze a number of dependent variables that I considered to be particularly susceptible to interviewer effects, namely (a) income, as a sensitive (and open-ended) factual question; (b) political participation, as a behavioral question; (c) political anomy, as an attitudinal question; and (d) healthy eating behavior, as an additional behavioral question.

Income was measured with the question: "How much money is at your disposal on average per month, during the current semester?"

Political participation was measured using a list of twelve political activities. The wording in the questionnaire was:

> If you wanted to have political influence or to make your point of view felt on an issue that was important to you: Which of the possibilities listed on these cards would you use? Which of them would you consider? Please name the corresponding letters.

In a previous study, I analyzed the effects of falsified data on the results of multivariate theory-driven OLS regression analyses, using the explanation of political participation as an example (Landrock, 2017). To investigate interviewer effects in the present study, the same dependent and independent variables were applied in a multilevel regression. Factor analysis revealed that that the factor *party-political activities* was an appropriate indicator for political participation. An additive index was calculated as a dependent variable measuring political participation. It consisted of the following three items:

- Participation in public discussions at meetings (factor loading: 0.701).
- Participation in a citizens' action group (factor loading 0.697).
- Voluntary work for a political party (factor loading 0.776).

Political anomy was measured with a scale consisting of four items that were summarized into an index that served as a third dependent variable (ZA & ZUMA, 2014). The items were:

- In spite of what some people say, the situation of the average man is getting worse, not better.
- It's hardly fair to bring a child into the world with the way things look for the future.
- Most public officials are not really interested in the problems of the average man.
- Most people don't really care what happens to the next fellow.

Healthy eating behavior was measured with the question: "On how many days per week do you eat healthy?" to analyze interviewer effects. I have used this variable in the past to explore the impact of falsifications on substantial findings in social science research on the basis of the theory of planned behavior (Landrock & Menold, 2016).

4.2 Independent Variables on the Individual Level

To implement multilevel regression models, statistically significant explanatory variables on the individual level were identified by estimating OLS regressions. These individual-level independent variables were included in the multilevel regression analyses presented in what follows. Given that my research interest here was to estimate interviewer effects, these variables may be considered as control variables.

For *income* as a dependent variable, the statistically significant explanatory variable on the respondent level – besides age – was the living situation, which was measured with the question: "Where are you living during the current semester?" This variable was dichotomized: The option "living with parents or relatives" was coded as 1; other options were coded as 0. The effect of age on income was positive.

Regarding the living situation, the analysis revealed that students who lived with their parents or relatives reported less income than students who did not.

For *political participation,* the statistically significant explanatory variables on the respondent level were internal political efficacy, political dissatisfaction, extremism (captured with the left–right scale), and (female) gender. The means of the individual items were calculated for both internal political efficacy and political dissatisfaction; all items were adapted from the ALLBUS 1998 questionnaire (see Koch et al., 1999).

The items used to measure *internal political efficacy* were:

- I would have the confidence to take on an active role in a group concerned with political issues.
- Politics is so complicated that somebody like me can't understand what's going on at all. (Reverse-scored item)

Political dissatisfaction was measured with the following three items:

- Only when differences in income and social status are large enough is there any incentive for personal achievement.
- Differences in social position between people are acceptable because they basically reflect what one has made of the chances one had.
- I consider the social differences in this country to be just on the whole.

To measure extremism, the left–right scale from the ALLBUS 1998 questionnaire was used:

> 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?

To operationalize extremism (see Lüdemann, 2001), the original 10-point rating scale (with the value 1 on the left end of the scale and the value 10 on the right end of the scale) was recoded in such a way that the original values between 1 and 10 were assigned the new values between 5 and -5. These new values were then squared, thereby yielding a measurement for extremism where the value 1 stands for a very small degree of extremism and the value 25 for a very high degree of extremism (integrating both the left and the right ends of the left–right scale). All of these variables, except extremism, were found to have significant positive effects in the real data. As extremism had a significant positive effect in the falsified data, this independent variable was nonetheless included in the analysis of interviewer effects (Landrock, 2017).

For the dependent variable *political anomy*, two statistically significant explanatory variables, economic dissatisfaction and external political efficacy were

identified. *Economic dissatisfaction* was measured with the question: "How would you generally rate the current economic situation in Germany?"

External political efficacy was measured with two items:

- Politicians don't care much about what people like me think. (Reverse-scored item)
- In general, politicians try to represent the people's interests.

Here, too, all items were adapted from the ALLBUS 1998 questionnaire. To operationalize external political efficacy, the means of the items were calculated (see Koch et al., 1999). Economic dissatisfaction was found to have a positive influence on political anomy, whereas external political efficacy had a negative effect.

To analyze interviewer effects on reported *healthy eating behavior*, a model based on the theory of planned behavior was adopted, which I applied in previous research on the impact of falsified data on substantive findings (Landrock & Menold, 2016).

The statistically significant independent variables for explaining healthy eating behavior on the individual level are the intention to eat healthily, perceived behavioral control, TV consumption, body mass index, doing sports, and preferring healthy desserts. The intention to eat healthily and perceived behavioral control were measured with two items each. These items were used to calculate an index for intention and for perceived behavioral control:

- In future I will eat healthy at least four days a week. (Intention)
- In the coming weeks I will eat healthy at least four days a week. (Intention)
- It is possible for me to eat healthy at least four days a week. (Perceived behavioral control)
- It is completely in my own hands to eat healthy at least four days a week. (Perceived behavioral control)

The questionnaire included the following question on TV consumption:

> Thinking about the days when you watch TV, how long on average do you watch TV on these days – I mean in hours and minutes?

Body mass index was calculated on the basis of the self-reported height and weight of respondents. The variable *doing sports* was dichotomized; respondents were asked to answer an open-ended question about which sports they took part in at least occasionally. A list of 12 desserts was used to find out whether the respondents preferred healthy desserts. The variable *preference for healthy desserts* was dichotomized. Healthy desserts (fruit curd, fruit salad, or yoghurt) were coded as 1; unhealthy desserts (mousse au chocolat, tiramisu, chocolate pudding, or pancakes) as 0.

As theory-driven explanatory variables, the intention to eat healthily and perceived behavioral control were found to have positive effects on reported healthy

eating behavior. TV consumption and body mass index had negative effects, whereas doing sports and preferring healthy desserts showed positive effects, at least in the falsified data.

4.3 Independent Variables on the Contextual Level

One aim of the present study was to identify interviewer characteristics on the contextual level that are linked to interviewer effects. The independent variables on the interviewer level that were tested are variables that are known to generally cause interviewer effects (see hypotheses in section 2 above). These variables are the *payment scheme* (payment per hour vs. payment per completed interview), the *interviewer's gender* and *income*, the *interviewer's response to the same question of the questionnaire*, and the *interviewer's experience*. Interviewers' personality traits were also tested, in particular *extraversion*, *conscientiousness*, and *perceived self-efficacy*, as they were considered relevant for analyzing interviewer effects.

First, the payment scheme was analyzed to determine whether the fact that an interviewer was paid per completed interview or per hour made a difference for the collected data, and therefore for the data quality. Winker et al. (2015) reported such an influence of the payment scheme on formal, non-content-related meta-indicators, for example non-differentiation. The payment scheme was varied in the research design: One half of the interviewers were paid per hour, the other half were paid per completed interview (see also section 3 above).

Many authors have described the core sociodemographic characteristics, namely gender, age, and education, as factors influencing interviewer effects (see Haunberger, 2006; Liu & Stainback, 2013). To my knowledge, researchers usually obtain only this basic information about interviewers from the fieldwork agencies, so that further interviewer characteristics typically cannot be analyzed. In the present study, I included the effects of the interviewers' gender as collected with the questionnaire completed by the interviewers themselves as respondents. Regarding age and education, the data show only small variances because all the interviewers were students and they were therefore very similar with respect to age and education. Instead, I considered the income of the interviewers, assuming that, in the case of the student population of interviewers, income would be an appropriate indicator for the socioeconomic background of an interviewer, which might lead to interviewer effects.

As mentioned above, the interviewers themselves also completed the survey questionnaire as respondents. Thus it was possible to include as an independent variable their responses to the same questions that the respondents were also asked. The interviewers' responses were included as an explanatory variable on the contextual level in order to test whether there were positive correlations between the respondents' answers and the interviewers' answers. Schanz (1981) reported posi-

tive correlations between the attitudinal and behavioral characteristics of interviewers and respondents.

A further relevant factor for the occurrence of interviewer effects is interviewer experience (Olson & Bilgen, 2011). The question used to measure this variable was whether the interviewer had ever conducted interviews before participating in the present study. The variable was dichotomized into interviewers with experience and interviewers without experience.

The questionnaire also included scales to measure the personality traits of the interviewers. To analyze the effects of the interviewers' personality traits on the respondents' responses, these traits were included in the multilevel analyses on the contextual level. Perceived self-efficacy was measured as agreement with the following three items (Beierlein, Kovaleva, Kemper, & Rammstedt, 2014) using a seven-point rating scale:

- I can rely on my own abilities in difficult situations.
- I am able to solve most problems on my own.
- I can usually solve even challenging and complex tasks well.

Afterwards, the means of the items were calculated.

To measure extraversion and conscientiousness, the ten-item Big Five Inventory (BFI-10; Rammstedt, Kemper, Klein, Beierlein, & Kovaleva, 2014) with a five-point rating scale was used:

I see myself as someone who...

- ...is reserved (Extraversion, reverse-scored item)
- ...is outgoing, sociable (Extraversion)
- ...tends to be lazy (Conscientiousness, reverse-scored item)
- ...does a thorough job (Conscientiousness)

For these variables, too, the means of each item were calculated.

4.4 Multilevel Regression Model

To test the hypotheses and to investigate whether the interviewers' characteristics influenced the respondents' answers (e.g., reported income), separate identical multilevel regression models were developed for the real and the falsified data. The statistical software Stata 12 was used to conduct the multilevel analyses. First, a null model without an independent variable and without the contextual level was estimated in order to assess the goodness of fit of the baseline model on the basis of log likelihood, or deviance (Hox, 1995). Second, to estimate interviewer-level variance the contextual level was included in the random-intercept-only model (RIOM) in order to be able to answer questions such as whether the income reported by the respondent depended on the interviewer – in other words, whether the incomes

of the respondents varied across interviewers. To this end, the intraclass correlation (ICC), which measures interviewer-level variance, was calculated. In the third step, the random-intercept model (RIM) was estimated. This model considers the influence of the individual respondent-level explanatory variables and controls for the contextual level. By including the interviewer-level explanatory variables of the contextual level (intercept-as-outcome model), direct effects of certain interviewer characteristics on respondents' responses were estimated. Thus, it could be determined, for example, whether the income reported by the respondents depended on the interviewers' gender. The results of the intercept-as-outcome model are shown in detail in Tables 4 and 5 (section 5.2).[5]

The likelihood-ratio test and McFadden's R-squared values were used to assess the goodness of fit of the model. With the likelihood-ratio tests, it was assessed, first, whether the multilevel approach was more appropriate than an OLS regression and, second, whether the estimated model extension (i.e., the reduction of deviance) was significant. McFadden's R-squared assesses model fit by comparing the log likelihood of the null model (i.e., the model without dependent variables and contextual level) with the log likelihood of the estimated model. According to Langer (2010, p. 756), values between 0.2 and 0.4 are excellent.

The dependent variables to be analyzed were required to be metric variables. Prior to the analyses, the independent variables were modified: The independent metric variables were grand-mean centered; the independent nominal variables were dichotomized and coded into binary variables.

5 Results

5.1 Interviewer Effects in Real Data

First, interviewer effects in the real data were analyzed. Table 2 shows the random-intercept-only model (RIOM) for all of the dependent variables.[6] The intraclass correlations varied between 0.017 and 0.067, which means that between 1.7 percent and 6.7 percent of the total variance is accounted for by the contextual level (i.e., the interviewer level). These interviewer effects are very small. Only healthy eating behavior, with an ICC of 0.067, showed slightly increased interviewer effects (see Groves & Magilavy, 1986; Mangione et al., 1992). The likelihood-ratio test measures the significance of the models and indicates whether a multilevel model

5 As an extension of the intercept-as-outcome models, the slope-as-outcome models were also estimated; they were not significant.

6 Regarding political anomy, it should be mentioned that there were a large number of missing values, due, in particular, to the item "Most public officials are not really interested in the problems of the average man" (56 missing values).

Table 2 Interviewer effects in the *real data* (random-intercept-only models, RIOMs)

	Dependent Variables			
RIOMs	Income σ^2 (SE)	Political Particip. σ^2 (SE)	Political Anomy σ^2 (SE)	Healthy Eating σ^2 (SE)
Resid. variance (respondents)	143206.6 (8553.957)	0.131 (0.007)	1.183 (0.071)	2.933 (0.165)
Resid. variance (interviewers)	3660.958 (3674.375)	0.002 (0.003)	0.063 (0.034)	0.210 (0.087)
ICC	0.025	0.017	0.050	0.067
LR test (p)	0.1356	0.1834	0.0114	0.0007
N	644	710	623	710

is more suitable than an OLS regression model. Regarding the dependent variables income and political participation, the RIOMs were not significant, which means that multilevel models were not appropriate and OLS regressions should be estimated instead. Regarding political anomy and healthy eating behavior, the RIOMs were significant; multilevel models could thus be preferred over OLS models. In the next step, the individual respondent-level variables were included in the model, and the random-intercept model (RIM) was developed. In the case of political anomy and healthy eating behavior as dependent variables, these models were not significant. Thus it can be assumed that interviewer effects scarcely exist in the real data.

5.2 Interviewer Effects in Falsified data

In the second step, interviewer effects in the falsified data were analyzed accordingly.[7] Table 3 shows the results of the RIOMs. The likelihood-ratio tests indicated that the models for all dependent variables were significant, which implies that the multilevel approach was more appropriate than the OLS regression approach. With values between 0.17 and 0.21, the intraclass correlations were much higher than in the real data, which means that the contextual level explained between 17 and 21% of the total variance. These strong interviewer effects indicate that individual characteristics, attitudes, and behaviors of the interviewers found their way into the

7 In the falsified data, there were a large number of missing cases in the case of income . I assume that the question is difficult to falsify and that the falsifiers therefore preferred to report item nonresponse.

Table 3 Interviewer effects in the *falsified data* (random-intercept-only models, RIOMs)

	Dependent Variables			
RIOMs	Income σ2 (SE)	Political Particip. σ2 (SE)	Political Anomy σ2 (SE)	Healthy Eating σ2 (SE)
Resid. variance (respondents)	30678.33 (1887.241)	0.102 (0.006)	1.125 (0.065)	1.869 (0.105)
Resid. variance (interviewers)	7913.874 (1964.437)	0.020 (0.005)	0.271 (0.065)	0.506 (0.115)
ICC	0.205	0.165	0.194	0.213
LR test (p)	0.0000	0.0000	0.0000	0.0000
N	606	708	681	710

falsified data. Thus, interviewer effects in the falsified data were further analyzed in order to determine which interviewer characteristics, attitudes or behaviors were particularly associated with interviewer effects.

In the third step, the RIOM was extended by including the respondent characteristics on the individual level (RIM, not shown here). Afterwards, the interviewer characteristics on the contextual level were included, thus developing the intercept-as-outcome model (IOM), which estimates the direct effects of the independent variables on the interviewer level. The further extensions of the IOM were not significant for any of the dependent variables. Therefore, the random-intercept, random-slope models with cross-level interactions could not be estimated. Table 4 shows the results of the final IOM for the dependent variables income and political participation.

As can be seen from Table 4, the models fit well: The likelihood-ratio test indicated that both the models themselves and the model extensions to IOMs were significant. The McFadden R-squared values of 0.16 and 0.64 were at least very reasonable.

The results show that all individual variables on the respondent level were significant, at least at the ten percent level, which is not surprising as they already proved to have significant influence in the previously performed OLS regressions. However, for the analysis of interviewer effects, the more relevant results were found on the contextual level. Significant effects on the dependent variables were not found for the payment scheme, the interviewers' personality traits, or the interviewers' experience. The interviewers' income had no significant effect on reported

political participation. However, for income and political participation as dependent variables, significant effects of the interviewers' gender and their answers to the same survey questions could be identified.

Table 4 Results of ML regression in the *falsified data* (intercept-as-outcome models, IOMs)

	Dependent Variables			
IOMs	Income		Polit. Particip.	
Fixed Part	Coeff.	SE	Coeff.	SE
Constant	725.907 ***	23.732	0.266 ***	0.036
Respondent level				
Age	10.381 ***	2.345	-	-
Living with parents/ relatives (ref.: no)	-176.879 ***	21.467	-	-
Internal political efficacy	-	-	0.128 ***	0.011
Political dissatisfaction	-	-	0.034 +	0.019
Gender (ref.: m)	-	-	0.035 +	0.019
Extremism	-	-	0.017 ***	0.003
Interviewer level				
Payment per hour (ref.: per int.)	2.435	23.428	-0.025	0.035
Gender (ref.: m)	-51.359 +	26.539	0.086 *	0.039
Income	-	-	0.000	0.000
Interviewer's answer	0.114 *	0.053	0.259 ***	0.052
Experience (ref.: no)	-4.696	29.644	-0.034	0.044
Extraversion	-1.050	14.651	0.017	0.022
Conscientiousness	17.575	15.002	0.022	0.022
Perceived self-efficacy	2.372	12.341	-0.013	0.019
Random Part	σ^2	SE	σ^2	SE
Respondents' residual variance	26933.240	1797.859	0.074	0.005
Interviewers' residual variance	4784.561	1509.125	0.010	0.003
Model fit				
Log likelihood	-3392.254		-92.393	
N	516		579	
LR test (p)	0.0000		0.0000	
LR test model extens. (p)	0.0000		0.0000	
McFadden's R^2	0.1641		0.6433	

Notes: *** p<0.001; ** p<0.01; * p<0.05; + p<0.10

Female falsifying interviewers tended to report lower incomes and higher values for political participation of the respondents than did male falsifying interviewers. Evidence was found that the gender of the interviewer tended to affect reported income and political participation in the case of the falsified data. It was also found that the interviewers' answers to the same questions had a positive effect on the reported respondents' answers. Thus, there were positive correlations between the falsifiers' attitudes and behaviors and the falsified reported attitudes and behaviors of the respondents. Presumably, the interviewers used their own income and political participation as a knowledge base for what a realistic income and political participation level might be for the interviews they were falsifying.

The models estimated for political anomy and healthy eating behavior as dependent variables yielded very similar results (Table 5). In both cases, the interviewers' answers to the same questions had a positive effect on the falsified reported answers of the respondents. In the case of healthy eating behavior as a dependent variable, the interviewers' gender affected the reported falsified response. Male falsifiers reported higher values for healthy eating. Thus, an impact of the attitudes and behaviors of the falsifying interviewers on all four analyzed variables could be identified.

Table 5 Results of ML regression in the *falsified data* (intercept-as-outcome models, IOMs)

	Dependent Variables			
IOMs	Polit. Anomy		Healthy Eating	
Fixed Part	Coeff.	SE	Coeff.	SE
Constant	1.691 ***	0.130	4.580 ***	0.140
Respondent level				
External political efficacy	-0.544 ***	0.045	-	-
Economic dissatisfaction	0.091	0.079	-	-
Intention	-	-	0.353 ***	0.032
Perceived behavioral control	-	-	0.359 ***	0.046
TV consumption	-	-	-0.003 **	0.001
Doing sports (ref.: no)	-	-	0.117 +	0.070
Preference for health desserts (ref.: no)	-	-	0.005	0.010
BMI	-	-	-0.100 ***	0.018
Interviewer level				
Payment per hour (ref.: per interview)	0.041	0.126	-0.089	0.133
Gender (ref.: m)	-0.229	0.148	-0.341 *	0.147
Income	0.000	0.000	0.000	0.000
Interviewer's answer	0.195 ***	0.054	0.160 ***	0.039
Experience (ref.: no)	-0.026	0.156	0.020	0.163
Extraversion	0.085	0.079	0.123	0.083
Conscientiousness	0.032	0.081	0.130	0.083
Perceived self-efficacy	-0.079	0.067	-0.079	0.070
Random Part	σ^2	SE	σ^2	SE
Respondents' resid. variance	0.896	0.057	0.998	0.063
Interviewers' resid. variance	0.133	0.042	0.143	0.047
Model fit				
Log likelihood	-797.383		-827.605	
N	565		565	
LR test (p)	0.0000		0.0000	
LR test model extension (p)	0.0000		0.0000	
McFadden's R^2	0.2613		0.3703	

Notes: *** p<0.001; ** p<0.01; * p<0.05; + p<0.10

5.3 Summary and Review of Hypotheses

First, I will review the two general hypotheses:

H1: Interviewer effects occur both in real and in falsified data.

This hypothesis cannot be confirmed. Interviewer effects were identified in the falsified data but not in the real data.

H2: The interviewer effects in falsified data are larger than in real data.

This hypothesis can be clearly confirmed. Large interviewer effects occurred in the falsified data, whereas interviewer effects could not be identified in the real data.

Next, I will review the more specific hypotheses regarding characteristics of the interviewers that may cause interviewer effects:

H3a: The core sociodemographic characteristics of the interviewers affect the reported responses.

As no effects of the core sociodemographic characteristics of the interviewers were measurable in the real data, this hypothesis must be rejected for the real data. With regard to the falsified data, the analysis of the effect of the interviewers' gender on the dependent variables revealed that female falsifiers reported lower income, higher political participation, and lower values for healthy eating behavior than did their male counterparts. The interviewers' age and education were too homogeneous to be tested. With the exception of income as a dependent variable (see H3c), the interviewers' income does not appear to have affected the falsified responses. Accordingly, for the falsified data, the hypothesis can be confirmed with respect to gender.

H3b: The magnitude of interviewer effects depends on the interviewer's experience.

This hypothesis could not be confirmed for the real or the falsified data: No effect of interviewer experience on any of the dependent variables was found.

H3c: Associations exist between the behaviors and attitudes of interviewers and the reported behaviors and attitudes of the respondents they interview.

This hypothesis cannot be confirmed for the real data, where no interviewer effects were found. However, strong evidence was found in support of the hypothesis in the falsified data: For all four dependent variables, significant positive correlations were found between the interviewers' answers as respondents and the falsified answers to the same survey questions.

H3d: The occurrence and magnitude of interviewer effects depends on the personality traits of the interviewer.

This hypothesis cannot be confirmed for the real data or for the falsified data. No effects of the personality traits on the dependent variables could be identified either in the real data or the falsified data.

H3e: The magnitude of interviewer effects depends on the interviewer payment scheme used (payment per completed interview vs. payment per hour).

This hypothesis cannot be confirmed for the real data or for the falsified data. Although previous research (see Winker et al., 2015) has shown that the payment scheme used (payment per completed interview vs. payment per hour) generally has an impact on the collected data, the present analyses did not detect effects of the payment scheme.

In summary, it can be stated that no interviewer effects of any kind were found in the real data. In the falsified data, the occurrence and magnitude of interviewer effects does not appear to have depended on the interviewers' experience or personality traits, or on the payment scheme used. However, effects of the interviewers' gender were found on the falsified reported income, political participation, and eating behavior of respondents. Furthermore, the interviewers' own attitudes and behaviors were correlated with the falsified reported attitudes and behaviors of the respondents. Thus, the falsifiers' attitudes and behaviors found their way into the falsified data and influenced the data reported as answers of the respondents.

6 Conclusions and Recommendations

The findings of the present study suggest that interviewer effects are clearly stronger in falsified data than in real data: The real data, derived from actual conducted interviews, does not appear to be contaminated by interviewer effects at all. This can be taken as an indication of high data quality, which may be due to the fact that the real interviews were audio-recorded and the fieldwork was intensively monitored. By contrast, very strong interviewer effects were measured in the falsified dataset. This suggests that the process of falsifying leads to a pronounced impact of the falsifiers' sociodemographic characteristics, attitudes, and behaviors on the data reported as answers of the respondents.

However, the interviewer effects (or, more precisely, "falsifier effects") identified in the falsified data were smaller than expected. One reason for this may be that both the respondents and the interviewers were students. Therefore, the falsifiers were familiar with the respondents' social reality and were able to give realistic answers – which reduced the magnitude of the interviewer effects. (This may also be a reason for the absence of interviewer effects in the real data.) A second reason why interviewer effects in the falsified data were smaller than expected may be that, despite the fact that the dependent variables used were empirically shown to

be susceptible to interviewer effects, more appropriate dependent variables could possibly have been found to analyze interviewer effects.

The fact that neither the payment scheme nor the interviewers' experience caused interviewer effects is surprising because current findings in the literature suggest that they should have. Winker et al. (2015) found that the payment scheme had an impact on formal, non-content-related meta-indicators such as non-differentiation. However, the present study analyzed content-related dependent variables. A further reason why the payment scheme did not have the hypothesized influence could be that the instructed falsifiers in the present experimental study had an intrinsic motivation to participate in the study and were therefore less frustrated by payment per completed interview than an interviewer in a real fieldwork setting might have been. Moreover, the interviewers in the present study selected the respondents on the university campus and interviewed them themselves. In a real fieldwork setting, the interviewers must contact certain predefined target persons, which may be time-consuming. In such a case it would appear plausible that the payment scheme would make a difference and that payment per hour might enhance motivation to contact the predefined target person. The lack of support for the hypothesized influence of interviewer experience might be due to the fact that the students who stated that they had conducted interviews before were still less experienced than the experienced interviewers in the studies in which interviewer effects have been found.

One limitation of the present study is the fact that the respondents and interviewers were students and that core sociodemographic characteristics, such as age and education, displayed only small variance. Moreover, in a real fieldwork setting, it would hardly be possible to implement an experimental approach such as that employed here. Nonetheless, I assume that the present results are generalizable, not least because interviewers in social science research and market research are often students. However, further research will be needed to confirm the generalizability of my results to real survey settings.

A number of recommendations can be derived from the present findings. First, researchers conducting interviewer-based surveys should collect as much information about the interviewers as possible and feasible (see Bogner & Landrock, 2016; Winker et al., 2015). In particular, as the present study shows, interviewer responses to the same questions that the respondents are asked are highly suitable for detecting interviewer effects in the case of falsified interviews. The interviewers could be requested to complete the survey questionnaire as part of interviewer training, for example. This would have at least two positive effects: First, the interviewers would familiarize themselves with the questionnaire, as a preparation for conducting the interviews; second, the researchers could get to know the interviewers.

A further recommendation that can be derived from the findings of the present study is that researchers using interviewer-based data should check the data for

interviewer effects, especially if they suspect that falsifications may have occurred. Falsification checking should be implemented at least by calculating intraclass correlations or conducting multilevel analyses as presented in this paper. This can be done for the entire dataset or only for suspicious cases – provided, of course, more than one interviewer is involved. If a large share of the variance is explained by interviewer-level variables, this may be an indication of contamination of the dataset by interviewer falsifications. In light of the fact that neither bivariate nor multivariate correlational analyses have proved effective in unambiguously establishing the existence of falsifications, the assessment strategies presented here may be very valuable for improving the quality and accuracy of survey data.

References

Beierlein, C., Kovaleva, A., Kemper, C. J., & Rammstedt, B. (2014). Allgemeine Selbstwirksamkeit Kurzskala (ASKU). *Zusammenstellung sozialwissenschaftlicher Items und Skalen*. doi:10.6102/zis35. Retrieved from http://zis.gesis.org/skala/Beierlein-Kovaleva-Kemper-Rammstedt-Allgemeine-Selbstwirksamkeit-Kurzskala-(ASKU)

Beullens, K., & Loosveldt, G. (2016). Interviewer effects in the European Social Survey. *Survey Research Methods*, *10*(2), 103-118. doi:10.18148/srm/2016.v10i2.6261.

Bogner, K., & Landrock, U. (2016). Response biases in standardised surveys. *GESIS Survey Guidelines*, Mannheim, Germany: GESIS – Leibniz Institute for the Social Sciences. doi: 10.15465/gesis-sg_en_016.

Groves, R. M., & Magilavy, L. J. (1986). Measuring and explaining interviewer effects in centralized telephone surveys. *Public Opinion Quarterly*, *50*(2), 251-266.

Haunberger, S. (2006). Das standardisierte Interview als soziale Interaktion: Interviewereffekte in der Umfrageforschung. *ZA-Information*, 58, 23-46.

Hox, J. J. (1995). *Applied multilevel analysis*. Amsterdam: TT-Publikaties.

Koch, A., Kurz, K., Mahr-George, H., & Wasmer, M. (1999). Konzeption und Durchführung der "Allgemeinen Bevölkerungsumfrage der Sozialwissenschaften" (ALLBUS) 1998. Mannheim: *ZUMA-Arbeitsbericht* 1999/02. Retrieved from http://nbn-resolving.de/urn:nbn:de:0168-ssoar-200413

Landrock, U. (2017). Explaining political participation: A comparison of real and falsified survey data. *Statistical Journal of the IAOS*, *33*(2), 447-458. doi: 10.3233/SJI-160270.

Landrock, U., & Menold, N. (2016). Validation of theoretical assumptions with real and falsified survey data. *Statistical Journal of the IAOS*, *32*(3), 305-312. doi: 10.3233/SJI-161020.

Langer, W. (2010). Mehrebenenanalyse mit Querschnittsdaten. In C. Wolf, & H. Best (Eds.), *Handbuch der sozialwissenschaftlichen Datenanalyse* (pp. 741-774), Wiesbaden: VS Verlag für Sozialwissenschaften.

Liu, M., & Stainback, K. (2013). Interviewer gender effects on survey responses to marriage related questions. *Public Opinion Quarterly*, *77*(2), 606-618.

Lüdemann, C. (2001). Politische Partizipation, Anreize und Ressourcen. Ein Test verschiedener Handlungsmodelle und Anschlusstheorien am ALLBUS 1998. In A. Koch, P. Schmidt, & M. Wasmer (Eds.), *Politische Partizipation in der Bundesrepublik*

Deutschland. Empirische Befunde und theoretische Erklärungen (pp. 43–71). Opladen: Leske and Budrich.

Mangione, T. W., Fowler, F. J., & Louis, T. A. (1992). Question characteristics and interviewer effects. *Journal of Official Statistics*, *8*(3), 293-307.

Menold N., & Kemper C. J. (2014). How do real and falsified data differ? Psychology of survey response as a source of falsification indicators in face-to-face surveys. *International Journal of Public Opinion Research*, *26*(1), 41-65. doi: 10.1093/ijpor/edt017.

Olson, K., & Bilgen, I. (2011). The role of interviewer experience on acquiescence. *Public Opinion Quarterly*, *75*(1), 99-114.

Rammstedt, B., Kemper, C. J., Klein, M. C., Beierlein, C., & Kovaleva, A. (2014). Big Five Inventory (BFI-10). *Zusammenstellung sozialwissenschaftlicher Items und Skalen*, doi:10.6102/zis76. Retrieved from http://zis.gesis.org/skala/Rammstedt-Kemper-Klein-Beierlein-Kovaleva-Big-Five-Inventory-(BFI-10)

Reuband, K.-H. (1990). Interviews, die keine sind. „Erfolge" und „Mißerfolge" beim Fälschen von Interviews. *Kölner Zeitschrift für Soziologie und Sozialpsychologie*, *42*(4), 706-733.

Schanz, V. (1981). Interviewereffekte: zusammenfassende Darstellung von Arbeiten, die im Rahmen zweier von ZUMA betreuter Projekte entstanden sind. *ZUMA Nachrichten 9*, 36-46.

Schnell, R. (1991). Der Einfluss gefälschter Interviews auf Survey-Ergebnisse. *Zeitschrift für Soziologie*, *20*(1), 25–35.

Schraepler, J.-P., & Wagner, G. G. (2003). Identification, characteristics and impact of faked interviews in surveys. An analysis by means of genuine fakes in the raw data of SOEP. *IZA Discussion Paper No. 969*, Forschungsinstitut zur Zukunft der Arbeit.

West, B. T., & Blom, A. G. (2016). Explaining interviewer effects: A research synthesis. *Journal of Survey Statistics and Methodology*, 0, 1-37. doi: 10.1093/jssam/smw024.

Winker P., Kruse, K.-W., Menold, N., & Landrock U. (2015). Interviewer effects in real and falsified interviews: Results from a large scale experiment. *Statistical Journal of the IAOS*, *31*(3), 423–434. doi: 10.3233/SJI-150908.

Zentralarchiv für empirische Sozialforschung (ZA) & Zentrum für Umfragen, Methoden und Analysen (ZUMA) e.V. (2014). Anomie (ALLBUS). *Zusammenstellung sozialwissenschaftlicher Items und Skalen*. doi:10.6102/zis58.Retrieved from http://zis.gesis.org/skala/ZA-ZUMA-Anomie-(ALLBUS)

methods, data, analyses | Vol. 11(2), 2017, pp. 189-206 DOI: 10.12758/mda.2017.04

Temporal Perspectives of Nonresponse During a Survey Design Phase

Taylor Lewis
U.S. Office of Personnel Management

Abstract

Invariably, full response is not achieved with a single survey solicitation, and so a sequence of follow-up attempts typically ensues in an effort to mitigate the potentially detrimental effects of nonresponse. Rather than permitting the follow-up campaign to continue indefinitely or until some preset response rate is met, a potentially more efficient alternative is to track a key point estimate in real-time as data is received and alter the survey design phase (i.e., modify the recruitment protocol) once the point estimate stabilizes. The notion of point estimate stability has been referred to as phase capacity in the survey methodology literature, and several methods to detect when it has occurred have been proposed in recent years. Noticeably absent from those works, however, is statistical theory providing insight into how point estimates can change during the course of data collection in the first place. The goal of this paper is to take a first step in developing that theory. To do so, the two established perspectives of survey nonresponse – deterministic and stochastic – are extended to account for the temporal dimension of responses obtained during a survey design phase. An illustration using data from the 2014 Federal Employee Viewpoint Survey is included to provide empirical support for the new theory introduced.

Keywords: responsive design, adaptive design, phase capacity, nonresponse bias, stopping rules

© The Author(s) 2017. This is an Open Access article distributed under the terms of the Creative Commons Attribution 3.0 License. Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.

1 Background

Unit nonresponse, which occurs whenever sampled cases (e.g., individuals, establishments) fail to respond to a survey request, is a ubiquitous problem faced by practitioners. Indeed, evidence abounds that response rates have been declining in surveys worldwide (Atrostic et al., 2001; de Leeuw & de Heer, 2002; Curtin et al., 2005; Brick & Williams, 2013). The typical data collection protocol in a survey involves making a sequence of follow-up attempts on cases yet to respond, which can take on a variety of forms depending on the survey's mode – reminder mailings, additional telephone calls, or revisits to a residence, to name a few. Each follow-up attempt generally yields more survey completes, which can be considered incoming *waves* of data. More follow-up attempts are ostensibly desirable, as they serve to reduce the nonresponse rate, but they can be costly and extend the field period, in turn delaying subsequent stages of the survey process, such as the reporting and analysis stages. And from a purely practical standpoint, empirical evidence (e.g., Table 1 in Potthoff et al., 1993; Table 1 in Lewis, 2017) suggests returns diminish with each subsequent wave; fewer and fewer completes are obtained, resulting in smaller and smaller changes in point estimates.

Rather than focusing on a target response rate or a predetermined number of completes, Groves & Heeringa (2006) advocate for the use of *responsive survey design*, which Schouten et al. (2013) note is a special case of *adaptive survey design* (Wagner, 2008). The premise of responsive survey design is to monitor in real-time the accumulating survey data in combination with data about the data collection process, referred to as *paradata* (Couper, 1998; Kreuter, 2013), to help inform decisions on whether, and when, to modify the current recruitment protocol. Groves & Heeringa (2006) define a *design phase* to be a data collection period with a stable sampling frame, sample, and recruitment protocol and *phase capacity* as the point during a design phase at which the additional responses cease influencing key estimates. Once phase capacity has been reached, some form of a design phase change is warranted. Examples include switching modes (de Leeuw, 2005), increasing the

Acknowledgements

An earlier version of this article appeared as part of the author's PhD dissertation "Testing for Phase Capacity in Surveys with Multiple Waves of Nonrespondent Follow-Up" from the Joint Program in Survey Methodology (JPSM) at the University of Maryland. The author would like to thank dissertation co-advisors Frauke Kreuter and Partha Lahiri for their encouragement, guidance, and feedback.

Disclaimer

The opinions, findings, and conclusions expressed in this article are those of the author and do not necessarily reflect those of the U.S. Office of Personnel Management.

Direct correspondence to

Taylor Lewis, U.S. Office of Personnel Management
E-mail: Taylor.Lewis@opm.gov

incentive offered (McPhee & Hastedt, 2012), or terminating nonrespondent follow-up altogether (Rao et al., 2008). While being an intriguing idea that could potentially lead to data collection efficiencies, an obstacle to those wishing to implement their approach was that no specific, calculable rule was given regarding how to formally test for phase capacity. The concept was only demonstrated visually in Figure 2 of their paper in which they plotted the trend of a key National Survey of Family Growth point estimate.

Over the last ten or so years, several phase capacity testing methods have emerged in the literature. The first was Rao et al. (2008), who developed a set of closely related methods to determine whether the most recent wave of data produced a statistically significant change in a sample mean. Lewis (2017) proposed a variant to their general approach amenable to any kind of point estimate, not strictly sample means. Wagner & Raghunathan (2010) took a prospective approach to testing for phase capacity, deriving a rule for determining whether or not a pending follow-up attempt was necessary. In addition, Moore et al. (2016) proposed identifying phase capacity based on coefficient of variation thresholds of an overall and unconditional partial *R-indicator* (Schouten et al., 2009; Schouten et al., 2012).

Noticeably absent in the works cited above is statistical theory to provide insight into the phenomenon of point estimate stability. That is, there is no theory offered to answer the following primordial question: How is it possible for a point estimate to change (or not change) over the course of a design phase? The works typically discuss the traditional nonresponse theory, but the traditional theory falls short because it is rooted in treating the act of responding as an all-or-nothing, yes-or-no event. In other words, the temporal dimension of the response process is not explicitly considered. This paper aims to fill that gap in the literature by extending the two traditional perspectives of nonresponse – deterministic and stochastic – to account for the timing of responses received during a survey design phase. Restricting the focus to a sample mean, we derive expressions of expected change to be observed with each new wave of responses obtained. These expressions are enlightening and provide a theoretical underpinning for the empirical tendency for point estimates computed from the accumulating data to deviate less, relatively speaking, later on in a survey design phase (e.g., Figure 3 in Peytchev et al. 2009; Figure 3 in Wagner, 2010; Figure 1 in Lewis, 2017).

The paper is structured as follows. In Section 2, we review the two traditional perspective of nonresponse. In Section 3, we factor into those perspectives a temporal dimension to account for changes that may be observed during a survey design phase. A brief illustration is given in Section 4 using data from the 2014 Federal Employee Viewpoint Survey. We conclude in Section 5 by suggesting avenues for further research.

2 Traditional Nonresponse Perspectives

The typical survey's data collection campaign commences by selecting a random sample of size *n* from a sampling frame constructed to represent all *N* units in a finite population. It has long been known from survey sampling theory that a randomly selected sample, even one of moderate size, can be used to form unbiased (or approximately unbiased) estimates of the attributes of the target population. The conundrum introduced by unit nonresponse is that, because only a portion of the sample is observed, unbiasedness properties are no longer guaranteed. Restricting analysis to the observed data without making any statistical adjustments may introduce nonresponse error (Groves, 1989), or a deviation from the quantity that would be computed had data been available for the full sample.

As discussed in Chapter 1 of Groves & Couper (1998), the magnitude of nonresponse error in a simple random sample of size *n* depends on both the statistic at hand and the degree of dissimilarity between the *r* observed cases and the *m* missing cases ($r + m = n$). To consider one example, suppose we were interested in estimating a finite population mean $\bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i$. We can formulate an unbiased estimate from the full sample by finding $\hat{\bar{y}}_n = \frac{1}{n}\sum_{i=1}^{n} y_i$. In the presence of unit nonresponse, however, we do not have all of the necessary information to compute this estimate. If we were to substitute $\hat{\bar{y}}_r = \frac{1}{r}\sum_{i=1}^{r} y_i$, the sample mean of the *r* observed cases, as the estimate of the finite population mean, the nonresponse error would be

$$NRerror(\hat{\bar{y}}_r) = \left(\frac{m}{n}\right)(\hat{\bar{y}}_r - \hat{\bar{y}}_m) \tag{1}$$

where $\hat{\bar{y}}_m = \frac{1}{m}\sum_{i=1}^{m} y_i$ represents the mean of the *m* missing cases. In other words, nonresponse error is the product of the nonresponse rate and the difference in means between the observed and missing cases. Note, however, that in the presence of an unequal probability of selection sample design where each sampled case has been assigned a base weight equaling the inverse of its selection probability, one would need to substitute base-weighted versions of the two sample means in equation 1. Additionally, one would need to replace the term *m*/*n* with the base-weighted nonresponse rate.

Nonresponse error in a sample mean can be partitioned further to account for two or more causes of nonresponse. For instance, a common differentiation is the portion attributable to noncontact versus explicit refusal given that contact has

been made (e.g., Lynn et al., 2002). To see this, suppose that the m nonrespondents in the sample are comprised of m_{nc} cases never contacted and m_{ref} cases who were reached but declined to participate in the survey ($r + m_{nc} + m_{ref} = n$). If we let $\hat{\bar{y}}_{nc}$ denote the mean of the m_{nc} cases never contacted and let $\hat{\bar{y}}_{ref}$ denote the mean of the m_{ref} cases refusing to participate, then the nonresponse error can be expressed as

$$NRerror(\hat{\bar{y}}_r) = \frac{m_{nc}}{n}(\hat{\bar{y}}_r - \hat{\bar{y}}_{nc}) + \frac{m_{ref}}{n}(\hat{\bar{y}}_r - \hat{\bar{y}}_{ref}) \tag{2}$$

Further decompositions of nonresponse error are possible, but the formulaic augmentation always abides by the same pattern: a new term is added representing the product of the prevalence of the group in the sample and the difference between the sample mean of the observed cases and the like for the group.

Lessler & Kalsbeek (1992) discuss at length the two traditional perspectives of nonresponse. The simpler view is the *deterministic* perspective, which stipulates that the N units on the sampling frame are comprised of two types: (1) a set of R units that will always respond when sampled; and (2) a set of M units that will never respond. Under this view, Valliant et al. (2013, equation 13.1) report that the nonresponse bias is

$$NRbias(\hat{\bar{y}}_r) = \left(\frac{M}{N}\right)(\bar{y}_R - \bar{y}_M) \tag{3}$$

where $\bar{y}_R$ represents the population mean of the units that always respond and $\bar{y}_M$ represents the like for units that never respond. Despite the resemblance to equation 1, equation 3 is expressed in terms of finite population quantities. In fact, the quantity in equation 1 can be considered an estimate of the quantity in equation 3.

An arguably more realistic view of nonresponse is the *stochastic* perspective, which assumes instead that all units in the finite population have some probability, or *propensity*, of responding to the survey request, a value between 0 and 1 frequently denoted ϕ_i. The concept and terminology are most often credited to Rosenbaum & Rubin (1983), but one can argue that the ideas trace back as far as Hartley (1946) and Politz & Simmons (1949). Given fixed propensities, if we let $\bar{\phi} = \frac{1}{N}\sum_{i=1}^{N}\phi_i$ symbolize the average response propensity for all N population units, Bethlehem (1988) showed that the nonresponse bias introduced by utilizing $\hat{\bar{y}}_r$, the sample mean for only the observed portion of the sample data, is approximately equal to

$$NRbias(\hat{\bar{y}}_r) \approx \frac{1}{N\bar{\phi}}\sum_{i=1}^{N}(\phi_i - \bar{\phi})(y_i - \bar{y}) \tag{4}$$

which reveals how the bias is proportional to the population covariance of the propensities and the survey outcome variable. A preliminary result of the proof is that the expected value of $\hat{\bar{y}}_r$ over the sampling and the nonresponse mechanisms is $\frac{\sum_{i=1}^{N} \phi_i y_i}{\sum_{i=1}^{N} \phi_i}$, which can be interpreted as the propensity-weighted mean of the outcome variable in the population. Derivations appearing in the next section will make use of that result.

The expression in equation 4 attributable to Bethlehem (1988) can be related to the three missingness mechanisms defined by Little & Rubin (2002). The first is that data are *missing completely at random* (MCAR), which is to say that all units in the population share the same propensity, or $\phi_i = \bar{\phi}$. In such a situation, there is no bias in $\hat{\bar{y}}_r$, because the first term in the summation is 0. The second mechanism, the one justifying most of the procedures used in practice to compensate for unit nonresponse, is that data are *missing at random* (MAR). Nonresponse adjustment techniques predicated on this mechanism exploit auxiliary data known for all sample units, both respondents and nonrespondents, such as information from the sampling frame or paradata. The MAR assumption permits response propensities to vary amongst sample units with different auxiliary variable profiles, but supposes that the propensities are identical for all sample units with the same profile. Hence, data are assumed MCAR conditional on the sample units' auxiliary variables. The third mechanism is the most perilous, data that are *not missing at random* (NMAR), meaning the sample units' response propensities vary as a function of the outcome variable beyond what can be explained (and adjusted for) by the auxiliary variables.

3 Alternative Nonresponse Perspectives to Frame the Phase Capacity Problem

The purpose of this section is to introduce extensions to the traditional nonresponse perspectives outlined in the previous section. These extensions are motivated by the objective of providing theoretical insight into how a sample mean can change, and eventually stabilize, over the course of a survey design phase. Both the deterministic and stochastic perspectives are considered.

A straightforward extension of the ideas behind the deterministic perspective of nonresponse for a survey collecting data over K waves is to conceptualize the N population units as falling within one of $K+1$ mutually exclusive and exhaustive domains: K domains of size $N_1, N_2, \ldots, N_K$ comprised of units that, if sampled, will

always respond to the survey during the k^{th} wave, and a domain of size M comprised of units that will never respond. Because of the empirical tendency for the number of respondents to decrease with each subsequent follow-up attempt within a survey design phase (e.g., Table 1 in Potthoff et al., 1993; Table 1 in Lewis, 2017), it seems reasonable to expect the N_k's to decrease in size as k increases.

Without loss of generality, as before, let us assume a simple random sample of size n has been selected and we are interested in making inferences on a finite population mean. We can expect the wave-specific respondent counts $r_1, r_2, \ldots, r_K$ and the count of nonrespondents m ($r_1 + r_2 + \ldots + r_K + m = n$) to fall approximately in proportion to their respective prevalences in the population – that is, $E(r_k) = n(N_k/N)$ for $k = 1, \ldots, K$ and $E(m) = n(M/N)$. Provided $r_k > 1$ for all K waves, we can express the ultimate respondent sample mean as $\hat{\bar{y}}_r = \sum_{k=1}^{K} \frac{r_k}{r} \hat{\bar{y}}_{r_k}$, where $r = \sum_{k=1}^{K} r_k$ and $\hat{\bar{y}}_{r_k}$ represents the sample mean of the r_k cases responding during wave k, specifically. Following the same strategy used to partition nonresponse error in equation 2, we can conceive of $\hat{\bar{y}}_1^k = \frac{\sum_{j=1}^{k} r_j \hat{\bar{y}}_{r_j}}{\sum_{j=1}^{k} r_j}$, the respondent mean using data from waves 1 to k inclusive ($k < K$) (i.e., calculated using data from the $r_1, r_2, \ldots, r_k$ respondents thus far obtained) as susceptible to nonresponse error due to the fact that there have been m nonrespondents drawn into the sample with mean $\hat{\bar{y}}_m$ that will never respond and $\sum_{k^*=k+1}^{K} r_{k^*}$ cases that have yet to respond:

$$NRerror(\hat{\bar{y}}_1^k) = \hat{\bar{y}}_1^k - \hat{\bar{y}}_n = \frac{m}{n}(\hat{\bar{y}}_1^k - \hat{\bar{y}}_m) + \sum_{k^*=k+1}^{K} \frac{r_{k^*}}{n}(\hat{\bar{y}}_1^k - \hat{\bar{y}}_{r_{k^*}}) \tag{5}$$

We can consider $\hat{\bar{y}}_1^1$ an estimate of $\bar{y}_1^1$, the mean of the population domain consisting of N_1 cases, and $\hat{\bar{y}}_1^2$ an estimate of $\bar{y}_1^2$, the mean of the population domain consisting of $N_1 + N_2$ cases, and so on. In terms of conventional statistical hypothesis testing, methods to test for phase capacity, at least those described in Rao et al. (2008) and Lewis (2017), use the accumulating data to assess H_0: $\delta_{k-1}^k = \bar{y}_1^{k-1} - \bar{y}_1^k = 0$ versus H_1: $\delta_{k-1}^k = \bar{y}_1^{k-1} - \bar{y}_1^k \neq 0$. Granted, the hypotheses can be written in terms of other population parameters, and non-zero differences for that matter.

Note, however, that the difference specified in the hypotheses above can be re-expressed as $\delta_{k-1}^k = (\bar{y}_1^{k-1} - \bar{y}_n) - (\bar{y}_1^k - \bar{y}_n)$, which reveals a parallel interpretation, and key finding, that testing for phase capacity is tantamount to testing whether there is any change with respect to nonresponse bias. In other words, if the cumu-

lative sample mean has not changed with the most recent wave of data set, then nonresponse bias has neither decreased nor increased.

The sample-based estimate of δ_{k-1}^{k} is $\hat{\delta}_{k-1}^{k} = \hat{\bar{y}}_1^{k-1} - \hat{\bar{y}}_1^{k}$, which can be re-expressed as follows:

$$
\begin{aligned}
\hat{\delta}_{k-1}^{k} &= \hat{\bar{y}}_1^{k-1} - \hat{\bar{y}}_1^{k} \\
&= (\hat{\bar{y}}_1^{k-1} - \hat{\bar{y}}_n) - (\hat{\bar{y}}_1^{k} - \hat{\bar{y}}_n) \\
&= NRerror(\hat{\bar{y}}_1^{k-1}) - NRerror(\hat{\bar{y}}_1^{k}) \\
&= \frac{m}{n}(\hat{\bar{y}}_1^{k-1} - \hat{\bar{y}}_m) + \sum_{k^*=k}^{K} \frac{r_{k^*}}{n}(\hat{\bar{y}}_1^{k-1} - \hat{\bar{y}}_{r_{k^*}}) - \frac{m}{n}(\hat{\bar{y}}_1^{k} - \hat{\bar{y}}_m) - \sum_{k^*=k+1}^{K} \frac{r_{k^*}}{n}(\hat{\bar{y}}_1^{k} - \hat{\bar{y}}_{r_{k^*}}) \\
&= \frac{m}{n}(\hat{\bar{y}}_1^{k-1} - \hat{\bar{y}}_m - \hat{\bar{y}}_1^{k} + \hat{\bar{y}}_m) + \frac{r_k}{n}(\hat{\bar{y}}_1^{k-1} - \hat{\bar{y}}_{r_k}) + \sum_{k^*=k+1}^{K} \left(\frac{r_{k^*}}{n}(\hat{\bar{y}}_1^{k-1} - \hat{\bar{y}}_{r_{k^*}} - \hat{\bar{y}}_1^{k} + \hat{\bar{y}}_{r_{k^*}}) \right) \\
&= \frac{m}{n}(\hat{\bar{y}}_1^{k-1} - \hat{\bar{y}}_1^{k}) + \sum_{k^*=k+1}^{K} \left(\frac{r_{k^*}}{n}(\hat{\bar{y}}_1^{k-1} - \hat{\bar{y}}_1^{k}) \right) + \frac{r_k}{n}(\hat{\bar{y}}_1^{k-1} - \hat{\bar{y}}_k) \\
&= \left(\frac{m + \sum_{k^*=k+1}^{K} r_{k^*}}{n} \right)(\hat{\bar{y}}_1^{k-1} - \hat{\bar{y}}_1^{k}) + \frac{r_k}{n}(\hat{\bar{y}}_1^{k-1} - \hat{\bar{y}}_k)
\end{aligned}
\tag{6}
$$

which illustrates how the observed change in the sample mean is equal to the sum of two terms: (1) the product of the portion of sample cases yet to be observed following wave k and the most recently observed change in the cumulative sample mean; and (2) the product of the portion of sample cases responding during wave k, specifically, and the difference between cumulative sample mean as of the previous wave and the sample mean of those responding during wave k. Because the r_k's tend to decrease as k increases, we would expect both terms to get closer and closer to zero. With respect to the first term, this is because $\hat{\bar{y}}_1^{k}$ consists of fewer and fewer new values relative to $\hat{\bar{y}}_1^{k-1}$, causing the difference $\hat{\bar{y}}_1^{k-1} - \hat{\bar{y}}_1^{k}$ to become smaller and smaller. With respect to the second term, this is because the multiplicative factor r_K/n gets progressively smaller.

We next consider augmentations with respect to the stochastic perspective of nonresponse. The fundamental difference is that we must broaden the idea of a single response propensity ϕ_i for the i^{th} population unit into a K-dimensional vector of wave-specific propensities, $\phi_i = [\phi_{1i}, \phi_{2i}, ..., \phi_{Ki}]$, where each entry represents the unit's propensity to respond during the k^{th} wave, specifically. This implies that the

response process for the i^{th} sample unit abides by a multinomial distribution with $K + 1$ events: responding during one of the K waves or not responding. Because all events are disjoint, we can treat the probability of responding by a particular wave as the sum of the entries in ϕ_i from the first position up to and including the entry indexing that wave. For example, the probability of the i^{th} sample unit responding before or during wave k is $\phi_{1i}^{k} = \sum_{j=1}^{k} \phi_{ji}$.

Alluded to earlier, a key preliminary result in the derivation of Bethlehem's (1988) nonresponse bias formula is that, given a set of fixed response propensities, the expectation of the sample mean from any sample design is shown to equal

$$E(\hat{\bar{y}}_r) = \frac{\sum_{i=1}^{N} \phi_i y_i}{\sum_{i=1}^{N} \phi_i} \tag{7}$$

which is a weighted mean for all population units, where the response propensity serves as the weight. Using this result, we can reason that the expectation of the sample mean at the first wave is $E(\hat{\bar{y}}_1^1) = \frac{\sum_{i=1}^{N} \phi_{1i} y_i}{\sum_{i=1}^{N} \phi_{1i}} = \frac{\sum_{i=1}^{N} \phi_{1i}^1 y_i}{\sum_{i=1}^{N} \phi_{1i}^1}$, and that the expectation of the sample mean at the second wave is $E(\hat{\bar{y}}_1^2) = \frac{\sum_{i=1}^{N} \phi_{1i}^2 y_i}{\sum_{i=1}^{N} \phi_{1i}^2}$, and so on. Therefore, we can express the expectation of the difference between two adjacent-wave sample means as

$$E(\hat{\bar{y}}_1^{k-1} - \hat{\bar{y}}_1^{k}) = \frac{\sum_{i=1}^{N} \phi_{1i}^{k-1} y_i}{\sum_{i=1}^{N} \phi_{1i}^{k-1}} - \frac{\sum_{i=1}^{N} \phi_{1i}^{k} y_i}{\sum_{i=1}^{N} \phi_{1i}^{k}} \tag{8}$$

This difference will only exactly equal zero if $\frac{\sum_{i=1}^{N} \phi_{1i}^{k-1} y_i}{\sum_{i=1}^{N} \phi_{1i}^{k-1}} = \frac{\sum_{i=1}^{N} \phi_{ki} y_i}{\sum_{i=1}^{N} \phi_{ki}}$, but as k

increases, the ϕ_{ki}'s decrease, rendering the component of $\sum_{i=1}^{N} \phi_{1i}^{k} y_i$ attributable to $\sum_{i=1}^{N} \phi_{ki} y_i$ to become smaller, and the same with the component of $\sum_{i=1}^{N} \phi_{1i}^{k}$ attributable to $\sum_{i=1}^{N} \phi_{ki}$. Hence, just as we could from the extended deterministic perspective, we can extract theoretical justification from equation 8 for the empirical tendency of point estimate differences to get progressively smaller during a survey design phase.

4 Illustration in the 2014 Federal Employee Viewpoint Survey

The purpose of this section is to provide an empirical illustration of the concepts and expressions presented in the previous section using data from the 2014 Federal Employee Viewpoint Survey (FEVS) (www.fedview.opm.gov). The FEVS, formerly known as the Federal Human Capital Survey (FHCS), was first launched in 2002 by the U.S. Office of Personnel Management (OPM). Initially administered biennially, the Web-based survey is now conducted yearly on a sample of full- or part-time, permanently employed civilian personnel of the U.S. federal government.

With few exceptions, the 2014 FEVS sampling frame was derived from a comprehensive personnel database managed by OPM known as the Statistical Data Mart of the Enterprise Human Resources Integration (EHRI-SDM). A total of 839,788 individuals from over 80 agencies were sampled as part of a single-stage stratified design, where strata were defined by the cross-classification of work unit and whether or not the employee was part of the Senior Executive Service (SES) or equivalent. The latter was done so that executives could be sampled with certainty, as they represent a rare population domain of analytic interest. The work-unit stratification ensured adequate numbers of employees appeared in the sample for all pre-identified agency subdivisions for which a separate report was desired. For agencies with exceedingly intricate reporting needs, a census was conducted. See U.S. Office of Personnel Management (2014) for more details about the FEVS sampling methodology.

The FEVS instrument consists of 84 work environment questions and 14 demographic questions. The work environment questions are predominantly attitudinal, capturing responses via a five-point Likert-type scale, such as one ranging from Strongly Agree to Strongly Disagree or Completely Satisfied to Completely Dissatisfied. Tests of statistical significance are typically performed after collapsing these categories into the dichotomy of a positive/non-positive response. The key

Table 1 2014 Federal Employee Viewpoint Survey Items Comprising the Global Satisfaction Index

Item Number	Wording
40	I recommend my organization as a good place to work.
69	Considering everything, how satisfied are you with your job?
70	Considering everything, how satisfied are you with your pay?
71	Considering everything, how satisfied are you with your organization?

estimate from each item thus reduces to the proportion (or percentage) of employees who react positively to the statement posed, what the FEVS administration team refers to as a "percent positive" statistic. For purposes of the present illustration, we restrict the focus to percent positive statistics for the four items comprising the Global Satisfaction Index (GSI). These items were purposefully chosen because they represent a cross-section of the typical satisfaction dimensions the FEVS is designed to capture. The wording for the four items is summarized in Table 1.

The 2014 FEVS was administered between April 29 and June 13, 2014. Participating agencies were given a choice of two possible start dates, April 29 or May 6. Each agency's field period spanned six work weeks. At survey close, 392,752 completes had been obtained, corresponding to an overall response rate of 47.4% per formula RR3 of the American Association for Public Opinion Research (AAPOR) (2016).

With respect to the responsive survey design terminology attributable to Groves & Heeringa (2006), the 2014 FEVS data collection protocol can be considered a single survey design phase. On the survey's launch date, an email invitation containing the website URL and log-in credentials was sent to sampled employees. Five reminder emails were sent to those who had yet to respond, in weekly increments thereafter. A final, sixth reminder was sent on Friday morning of the sixth field period week with messaging emphasizing that the survey would close at the end of the day. In all, seven email notifications were sent. A natural demarcation of a data collection wave, the one used in this illustration, is the set of responses obtained between two chronologically adjacent email notifications.

Table 2 summarizes the wave-specific respondent counts for one example agency participating in 2014 FEVS that conducted a census of its $N = 5{,}188$ employees. The greatest number of responses was obtained in the first wave, followed by the second wave, with returns diminishing in subsequent waves. A total of $m = 1{,}592$ employees never responded, even after being sent seven email notifications. Though not shown here, comparable patterns hold for most other participating agencies. Thinking back to the second term of equation 6, this lends empirical

Table 2 Wave-Specific Response Distribution for an Example Agency Participating in the 2014 Federal Employee Viewpoint Survey

Data Collection Wave k	Count r_k	Percent of Sample $(r_k / n) * 100$
1	1,390	26.8
2	873	16.8
3	240	4.6
4	392	7.6
5	246	4.7
6	260	5.0
7	195	3.8
Nonrespondents	1,592	30.7
Total	5,188	100.0

credence to the assertion of the r_k terms decreasing as k increases, a major factor in the stabilization of a sample mean over the course of a survey design phase.

The decreasing r_k's also factor implicitly into the first product in equation 6, as is evident from Figure 1, which plots the trends in the cumulative sample means of the four GSI items using responses obtained through the given wave (i.e., the $\hat{\bar{y}}_1^k$'s) for the example agency. The cumulative means tend to increase with each new wave of data, at least for the early waves, but then stabilize around wave 5. The increasing pattern is an indication that the early responders are less positive than later responders, something Sigman et al. (2014) noted was widespread amongst agencies participating in the 2011 FEVS.

With respect to the stochastic perspective of nonresponse, recall the primary takeaway argument from equation 8 was that, because the wave-specific propensities (i.e., the ϕ_{ki}'s) tend to decrease as k increases, the component of $\sum_{i=1}^{N} \phi_{1i}^k$ attributable to $\sum_{i=1}^{N} \phi_{ki}$ and the component of $\sum_{i=1}^{N} \phi_{1i}^k y_i$ attributable to $\sum_{i=1}^{N} \phi_{ki} y_i$ should both become progressively smaller over the course of a design phase. When those respective components of the summations become negligible, phase capacity results.

To illustrate how this can happen, we can exploit information from the 2014 FEVS sampling frame. Specifically, using auxiliary information known for the entire population of $N = 5{,}188$ individuals in the agency, we utilized the employee's age, gender, an indicator of being a supervisor/non-supervisor, an indicator of being minority/non-minority race or ethnicity, and an indicator of working in the headquarters or field office, to fit a multinomial logistic regression model where the outcome variable was one of 8 possible events: responding during wave 1, 2, …, 7,

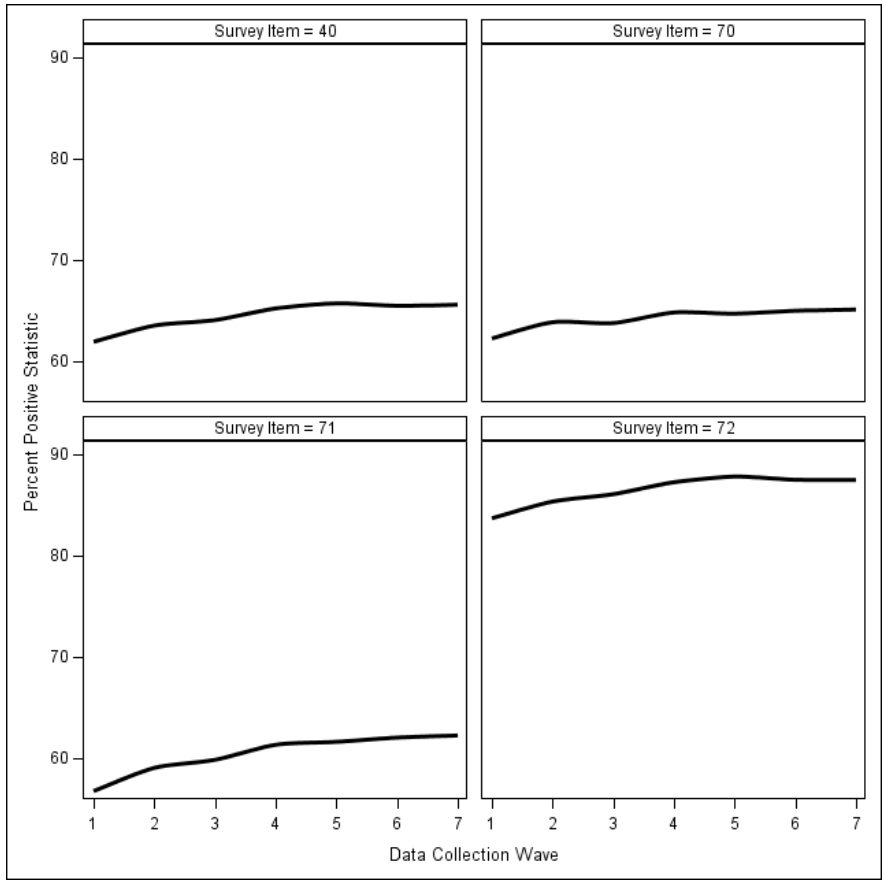


Figure 1 Trends in the Percent Positive Statistics for Items Comprising the Global Satisfaction Index over an Example Agency's 2014 Federal Employee Viewpoint Survey Data Collection Period

or not responding at all. This model was used to generate estimated wave-specific propensities, or $\hat{\phi}_{ki}$'s, which can serve as substitutes for the ϕ_{ki}'s.

Table 3 reports the proportions $\sum_{i=1}^{N} \hat{\phi}_{1i} / \sum_{i=1}^{N} \hat{\phi}_{1i}^{k}$ and $\sum_{i=1}^{N} \hat{\phi}_{1i} y_{ki} / \sum_{i=1}^{N} \hat{\phi}_{1i}^{k} y_{ki}$ for the four GSI items, where y_{ki} is an indicator variable equaling 1 for a positive response to a given item and 0 otherwise. Each of these proportions converges towards zero, which is to say that both the numerator and denominator terms of the expected value of the cumulative sample mean (see equation 8) change less and less. By wave 5, the proportional change is less than 10%, suggesting an ineffectual impact, which coincides with the point estimate stabilization observed in Figure 1.

Table 3 Proportions of Estimated Wave-Specific Response Propensities, and Proportions of the Products of Estimated Wave-Specific Response Propensities with GSI Positive/Non-Positive Indicator Variables for an Example Agency Participating in the 2014 Federal Employee Viewpoint Survey

Data Collection Wave k	Propensities	Propensities x Item 40	Propensities x Item 70	Propensities x Item 71	Propensities x Item 72
1	1.00	1.00	1.00	1.00	1.00
2	0.39	0.39	0.38	0.39	0.39
3	0.10	0.10	0.10	0.10	0.10
4	0.14	0.14	0.13	0.14	0.14
5	0.08	0.08	0.08	0.08	0.08
6	0.08	0.08	0.07	0.08	0.08
7	0.05	0.05	0.05	0.05	0.05

5 Discussion

Faced with downward pressures on response rates, practitioners must nowadays explore alternative strategies to more effectively and efficiently manage a survey's data collection process. One intuitive method for doing so is to monitor a key point estimate from the survey in real-time as completes are obtained and take note of when it stabilizes. This is the notion of phase capacity, as defined by Groves & Heeringa (2006), who argue that additional follow-up efforts are liable to be equally inefficacious. Instead, some form of change in the data collection protocol is warranted. In their terminology, a new design phase is in order.

Groves & Heeringa (2006) did not offer a formal method to test for phase capacity, but several techniques have since been proposed in the literature (Rao et al., 2008; Wagner & Raghunathan, 2010; Moore et al., 2016; Lewis, 2017). An important piece missing from those proposals, however, is statistical theory illuminating how (or when) point estimate changes could occur in the first place. The objective of this paper was to fill that void in the literature. Using the finite population mean as an example, we extended the traditional deterministic and stochastic perspectives of nonresponse to derive expressions of change that explicitly account for incoming waves of responses within a single design phase. To connect these ideas to practice and to secure empirical support of certain assumptions and assertions made during the derivations, we included an illustration using data from the

2014 Federal Employee Viewpoint Survey. In particular, focusing on four survey items for one example agency, we showed how the stabilization occurring around the fifth wave of data received was largely a function of the decreasing respondent counts (i.e., the r_k's in equation 6) and the associated decreasing (estimated) wave-specific propensities that factor into the two quotients in equation 8.

Of course, this paper is not without limitations. The first limitation is that we focused solely on a sample mean. Brick & Jones (2008) derive expressions of nonresponse bias for several other statistics. Modifications to those expressions accounting for the temporal dimension of nonresponse could have proven equally as enlightening. A second limitation is that, for tractability, the derivations presented in Sections 2 and 3 assumed no nonresponse adjustments have been made. In fact, the phase capacity testing methods proposed in Rao et al. (2008), Wagner & Raghunathan (2010), and Lewis (2017) call for nonresponse adjustments to be made prior to assessing whether point estimate stability has occurred. A third limitation is that the 2014 FEVS illustration only involved analysis of four survey items for one example agency. Although we argued that the patterns observed are very typical for the FEVS, both in terms of other items' percent positive statistics and other participating agencies, it is certainly conceivable that a comparable illustration within the design phase(s) of another survey could yield results less harmonious with the nonresponse theory extended in this paper.

Aside from addressing the limitations just cited, further research could extend the theory to account for two or more design phases within the same survey, two or more key outcome variables, or both. Another potential extension, motivated by findings in Olson & Groves (2012), would be to relax the assumption of fixed response propensities under the stochastic perspective of nonresponse, instead allowing them to vary in some way over the course of data collection. Finally, future research could investigate whether information gleaned from, say, estimated wave-specific response propensities could be carried forward in a meaningful way in an adaptive survey design approach (Schouten et al., 2013). For example, in the FEVS there are numerous agencies that conduct a perennial census. It seems foreseeable that prior survey response patterns, perhaps in combination with imputation or auxiliary information from the sampling frame, such as a variable highly correlated with one or more key outcome variables, could be used to derive measures similar in spirit to those derived in this paper to help support (or refute) evidence of phase capacity.

References

American Association for Public Opinion Research (AAPOR). (2016). *Standard definitions: Final dispositions of case codes and outcome rates for surveys* (9th edition). AAPOR.

Atrostic, B., Bates, N., Burt, G., & Silberstein, A. (2001). Nonresponse in US government household surveys: Consistent measures, recent trends, and new insights. *Journal of Official Statistics*, 17 (2), 209-226.

Bethlehem, J. (1988). Reduction of nonresponse bias through regression estimation. *Journal of Official Statistics*, 4 (3), 251-260.

Brick, M., & Jones, M. (2008). Propensity to respond and nonresponse bias. *Metron-International Journal of Statistics*, 66 (1), 51-73.

Brick, M., & Williams, D. (2013). Explaining rising nonresponse rates in cross-sectional surveys. *ANNALS of the American Academy of Political and Social Science*, 645 (1), 36-59.

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

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

de Leeuw, E., & de Heer, W. (2002). Trends in household survey nonresponse: A longitudinal and international comparison. In R. Groves, D. Dillman, J. Eltinge, & R. Little (Eds.), *Survey Nonresponse*, New York: Wiley.

de Leeuw, E. (2005). To mix or not to mix data collection modes in surveys. *Journal of Official Statistics*, 21 (2), 233-255.

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

Groves, R., & Couper, M. (1998). *Nonresponse in household interview surveys*. New York: Wiley.

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

Hartley, H. (1946). Discussion of "A review of recent statistical developments in sampling and sampling surveys" by Yates, F. *Journal of the Royal Statistical Society: Series A*, 109 (1), pp. 37-38.

Kreuter, F. (2013). *Improving surveys with paradata: Analytic uses of process information*. Hoboken: Wiley.

Lessler, J., & Kalsbeek, W. (1992). *Nonsampling error in surveys*. New York: Wiley.

Lewis, T. (2017). Univariate Tests for Phase Capacity: Tools for Identifying When to Modify a Survey's Data Collection Protocol. *Journal of Official Statistics* (in press).

Little, R., & Rubin, D. (2002). *Statistical analysis with missing data* (2nd edition). New York: Wiley.

Lynn, P., Clarke, P., Martin, J., & Sturgis, P. (2002). The effects of extended interviewer effort on nonresponse bias. In R. Groves, D. Dillman, J. Eltinge, & R. Little (Eds.), *Survey Nonresponse*, New York: Wiley.

McPhee, C., & Hastedt, S. (2012). More money? The impact of larger incentives on response rates in a two-phase mail survey. *Paper presented at the Federal Committee on Statistical Methodology (FCSM) Research Conference*.

Moore, J., Durrant, G., & Smith, P. (2016). Data set representativeness during data collection in three UK social surveys: Generalizability and the effects of auxiliary covariate choice. *Journal of the Royal Statistics Society: Series A*, online first edition.

Olson, K., & Groves, R. (2012). An examination of within-person variation in response propensity over the data collection field period. *Journal of Official Statistics*, 28 (1), 29-51.

Peytchev, A., Baxter, R., & Carley-Baxter, L. (2009). Not all survey effort is equal: Reduction of nonresponse bias and nonresponse error. *Public Opinion Quarterly*, 73 (4), 785-806.

Politz, A., & Simmons, W. (1949). An attempt to get the not-at-homes into the sample without callbacks. *Journal of the American Statistical Association*, 44 (245), 9-31.

Potthoff, R., Manton, K., & Woodbury, M. (1993). Correcting for nonavailability bias in surveys by weighting based on the number of callbacks. *Journal of the American Statistical Association*, 88 (424), 1197-1207.

Rao, R., Glickman, M., & Glynn, R. (2008). Stopping rules for surveys with multiple waves of nonrespondent follow-up. *Statistics in Medicine*, 27 (12), 2196-2213.

Rosenbaum, P., & Rubin, D. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70 (1), 41-55.

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

Schouten, B., Bethlehem, J., Beullens, K., Kleven, Ø., Loosveldt, G., Luiten, A., Rutar, K., Shlomo, N. & Skinner, C. (2012). Evaluating, comparing, monitoring, and improving representativeness of survey response through R-indicators and partial R-indicators. *International Statistical Review*, 80 (3), 382-399.

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

Sigman, R., Lewis, T., Yount Dyer, N., & Lee, K. (2014). Does the length of fielding period matter? Examining response scores of early versus late responders. *Journal of Official Statistics*, 30 (4), 651-674.

United States Office of Personnel Management. (2014). *Federal employee viewpoint survey results: Technical report*. Retrieved September 13, 2016 from the Federal Employee Viewpoint Survey website: http://www.fedview.opm.gov/2014/published/.

Valliant, R., Dever, J., & Kreuter, F. (2013). *Practical tools for designing and weighting survey samples*. New York: Springer.

Wagner, J. (2008). Adaptive survey design to reduce nonresponse bias. Ph.D. thesis, University of Michigan.

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

Wagner, J., & Raghunathan, T. (2010). A new stopping rule for surveys. *Statistics in Medicine*, 29 (9), 1014-1024.

Information for Authors

Methods, data, analyses (mda) publishes research on all questions important to quantitative methods, with a special emphasis on survey methodology. In spite of this focus we welcome contributions on other methodological aspects.

Manuscripts that have already been published elsewhere or are simultaneously submitted to other journals will not be considered. As a rule we do not restrict authors' rights. All rights remain with the author, and articles in mda are published under the CC-BY open-access license.

Mda aims for a quick peer-review process. All papers submitted to mda will first be screened by the editors for general suitability and then double-blindly reviewed by at least two reviewers. The decision on publication is made by the editors based on the reviews. The editorial team will contact the authors by email with the result at the latest eight weeks after submission; if the reviews have not been received by then, we provide a status update with a new target date.

When preparing a paper for submission, please consider the following guidelines:

- Please submit your manuscript via www.mda.gesis.org.
- The total length of the manuscript shall not exceed 10.000 words.
- Manuscripts should…
  - be written in English, using American English spelling. Please use correct grammar and punctuation. Non-native English speakers should consider a professional language editing prior to publication.
  - be typed in a 12 pt Roman font, double-spaced throughout.
  - be sent as MS Word documents.
  - start with a cover page containing the title of the paper and contact details / affiliations of the authors, but be anonymized for review otherwise.
- Please also send us an abstract of your paper (approx. 200 words), a brief biographical note (no longer than 250 words), and a list of 5-7 keywords for your paper.
- Acceptable formats for Graphics are
  - tiff
  - jpg (uncompressed, high quality)
  - pdf
- Please ensure a resolution of at least 300 dpi and take care to send hiqh-quality graphics. Line art images should have a resolution of 500-1000 dpi. Please note that we cannot print color images.
- The type area of our journal is 11.5 cm (width) x 18.5 cm (height). Please consider this when producing tables or graphics.
- Footnotes should be used sparingly.
- By submitting a paper to mda the authors agree to make data and program routines available for purposes of replication.

Please follow the APA guidelines when preparing in-text references and the list of references.

Entire Book:

Groves, R. M., & Couper, M. P. (1998). *Nonresponse in household interview surveys*. New York: John Wiley & Sons.

Journal Article (with DOI):

Klimoski, R., & Palmer, S. (1993). The ADA and the hiring process in organizations. *Consulting Psychology Journal: Practice and Research*, 45(2), 10-36. doi:10.1037/1061-4087.45.2.10

Journal Article (without DOI):

Abraham, K. G., Helms, S., & Presser, S. (2009). How social processes distort measurement: The impact of survey nonresponse on estimates of volunteer work in the United States. *American Journal of Sociology*, 114(4), 1129-1165.

Chapter in an Edited Book:

Dixon, J., & Tucker, C. (2010). Survey nonresponse. In P. V. Marsden & J. D. Wright (Eds.), *Handbook of Survey Research*. Second Edition (pp. 593-630). Bingley: Emerald.

Internet Source (without DOI):

Lewis, O., & Redish, L. (2011). *Native American tribes of Wisconsin*. Retrieved April 19, 2012, from the Native Languages of the Americas website: www.native-languages.org/wisconsin.htm

For more information, please consult the Publication Manual of the American Psychological Association (Sixth ed.).

gesis
Leibniz Institute for the Social Sciences

ISSN 1864-6956 (Print)
ISSN 2190-4936 (Online)
© of the compilation GESIS, Mannheim, July 2017