Unintended Behavioral Consequences Due to Repeated Measurement

Inauguraldissertation
zur Erlangung des akademischen Grades
eines Doktors der Sozialwissenschaften
der Universität Mannheim

Vorgelegt von

Ruben Lukas Bach

Hauptamtlicher Dekan der Fakultät für Sozialwissenschaften:

Prof. Dr. Michael Diehl

Erstbetreuerin:

Prof. Dr. Frauke Kreuter

Zweitbetreuer:

Prof. Joseph Sakshaug, PhD

Erstgutachter:

Prof. Joseph Sakshaug, PhD

Zweitgutachter:

Prof. Dr. Florian Keusch

Tag der Disputation:

14. März 2018

Acknowledgements

Many people have supported me during the work on this dissertation. First of all, I wish to express my deep gratitude to Stephanie Eckman, who has been an invaluable mentor for the last three years. She got me working in survey research in the first place, co-authored three papers of this dissertation, spent endless hours reviewing my work, and, although she left the IAB a few months after I started with this project, did an extraordinary job mentoring and supporting me from far away. I am also very grateful to my supervisor, new and old boss, Frauke Kreuter, who encouraged me many times to think a little bigger and reach a little further. Without Stephanie and Frauke's guidance, this project would not have been possible. Danke!

I also thank Joe Sakshaug and Florian Keusch for reviewing my thesis; my wonderful colleagues at the IAB and the fellows from the GradAB program for all the helpful discussions and for making the last three years more fun; the IAB GradAB for the financial support; the members of FK^{2}RG for the numerous discussions of my papers; Jessica Daikeler for co-authoring one of the papers; Fred Conrad for inviting me to spend three very productive months at the University of Michigan-Ann Arbor, and members of the MPSM Meth Lab for having me in their group!

Contents

List of Tables

List of Figures

Chapter 1

A Short Introduction and Summary

1.1 General remarks

Empirical research in the social sciences often requires the collection of information about research objects. When these objects are humans, researchers often use surveys and simply ask respondents questions to collect the information desired. Such information may be sociodemographic (e.g., the age or gender of the respondents) or it may refer to behavior shown by the respondents or attitudes held by them.

Collecting information about behaviors or attitudes often requires asking *similar* or even the *same* questions, sometimes even on multiple occasions, e.g., in longitudinal or panel surveys. The reasons for asking similar or identical questions either at a single occasion or at multiple ones are numerous. For example, researchers may be interested in a set of behaviors and attitudes that they believe occur together or where one behavior explains the occurrence of another. Similarly, researchers may want to gather the same details about several similar behaviors or attitudes (for example, motivations or consequences of a behavior) to test a theoretical model; they may be interested in studying changes in a behavior or an attitude over time or they may want to infer a causal relationship between changes in a behavior and changes in the respondent's environment from longitudinal data. As a result, respondents are often presented with questionnaires containing similar or identical elements or they may be presented with the same questionnaire on multiple occasions. It is even possible that researchers interview respondents with a questionnaire that contains similar questions *and* that these questions are asked on multiple occasions, e.g., in a panel survey.

Retrieving information from humans in surveys, however, may affect respondents in certain ways, just as the temperature of a thermometer may change the water's temperature it is supposed to measure. E.g., responding to questions in a survey may influence respondents' future behavior or it may influence the way how respondents report in the future. I refer to such effects of answering questions in a survey on respondents' future behavior and/or future reporting as *conditioning* effects. When conditioning affects respondents' behavior, I refer to it as *changes-in-behavior* conditioning. When survey participation conditions respondents' reporting, however, I refer to it as *changes-in-reporting* conditioning.

Furthermore, I refer to changes-in-reporting conditioning that occurs due to answering a set of similar questions within a single survey by *survey conditioning*, following Duan et al. (2007).[1] When respondents' behavior or reporting is conditioned by participating in several waves of the same survey, i.e., through participation in a panel survey, I refer to these forms of conditioning by *panel conditioning*.

Providing a methodological framework for the study of these conditioning effects and analyzing whether some of their forms actually occur in social science (longitudinal) survey data is the goal of this dissertation. That is, I study whether repeated measurement with similar or even the same questions in (panel) surveys leads to changes in the ways how respondents behave and/or report their behavior. Before explaining the need for the studies that form this dissertation, for clarification, I give examples of the different types of conditioning effects and I briefly review the consequences they may have for survey data first.

[1] There is no changes-in-behavior survey conditioning because survey conditioning occurs *within* a single survey and there is therefore simply no time for respondents to change their behavior. However, it is possible that respondents' behavior changes *after* participating in a single survey. Without measuring the same behavior for a second time, i.e., through a second interview, changes in that behavior would, however, not be uncovered.

1.2 Examples of the various forms of conditioning

(Repeatedly) answering a question about some form of behavior may lead to changes in this behavior (changes-in-behavior panel conditioning) (Waterton and Lievesley 1989). For instance, answering questions may stimulate respondents to think about the behavior under study and eventually even lead them to behave differently in the future. When respondents are interviewed for a second time with the same questionnaire (e.g., because researchers are interested in changes in that behavior over time or a causal effect of some intervention), respondents would then report on their adjusted behavior. However, had they not been interviewed before, their behavior would not have changed. In this way, respondents' behavior is conditioned by repeated panel survey participation.

Consider, for example, a survey that asks unemployed people about their job search strategies. Answering to the questions of the survey might stimulate an unemployed respondent to reflect on her search strategies. As a result of this cognitive process, she may start to reconsider her current search strategies and eventually adjust her strategies to others thought to be more successful. Similarly, it is also possible that an unemployed respondent who does not spend much time on job search realizes a conflict between her behavior and society's norms. She might realize that the socially desirable behavior would be to spend more time on job search in order to become employed again and contribute to societal wealth. As a result, she may then increase the time she spends on job search to bring her actual behavior in line with society's norms. In both cases, had she not participated in the survey, she would not have adjusted her job search behavior. Therefore, in this example, answering questions conditions the respondent's job search behavior.

Being surveyed, however, can also change the way a respondent answers survey questions (changes-in-reporting conditioning) (Waterton and Lievesley 1989). For example, a respondent may recall having answered similar or the same questions before. When she is interviewed with a second similar set of questions (survey conditioning) or with the same questions for the second time (panel conditioning), having already answered them in the past may influence how she answers the questions in the present. Had she not been interviewed before, her reporting would not have changed. In this way, the respondent's reporting behavior is conditioned by (repeated panel) survey participation.

Consider, for example, a survey that asks a series of simple "Yes"/"No" questions to determine the eligibility of a respondent for follow-up questions. If the respondent is eligible, she receives some extra questions before continuing with the remaining eligibility questions. If she is not eligible, she receives the next eligibility question right away. Such questions may, for example, ask for purchase behavior of certain goods. If the respondent confirms the purchase of the goods, follow-up questions could ask details about the purchase or the purchased goods themselves (as used in the Consumer Expenditure Survey, for example). Some respondents, however, may realize after a few eligibility and follow-up questions that saying "Yes" to the eligibility question results in additional questions. As a result, they may begin to say "No" to the eligibility questions to avoid the follow-up questions and thereby shorten the duration of the interview. In this way, some respondents may become worse reporters over time due to changes-in-reporting survey conditioning, because they report less accurately to later questions in a survey.[2]

However, it is also possible that respondents become better reporters over time due to changes-in-reporting conditioning. For example, some respondents of a survey that asks sensitive questions (e.g., about substance abuse) may feel uncomfortable revealing their behavior to the interviewer because the respondents perceive their own behavior as socially undesirable. In a second or third wave of the same survey, however, when respondents have been interviewed a few times, they may feel more comfortable revealing their actual behavior to the interviewer. One reason for this may be increased trust towards the interviewer due to familiarity. Thus, some respondents may become better reporters, i.e., report more accurately and honestly, over time due to changes-in-reporting conditioning.[3]

All of these examples show that respondents' behavior and/or their reporting *can* become conditioned by answering similar or the same questions, either within a single survey due to

[2] The example can also be extended to the panel survey case. If respondents remember the structure of the survey from one wave to another, they may become worse reporters not only within a single survey, but also over the waves of a panel survey.

[3] Respondents within a single survey may also become better reporters over time if, for example, they get a better understanding of the survey questions after having answered a few of them.

survey conditioning or over multiple waves of the same survey due to panel conditioning. In this dissertation, I test whether such effects occur in the surveys considered below. Before summarizing the four papers that form the core of my dissertation, I will briefly discuss why researchers should pay attention to these effects.

1.3 Consequences of conditioning effects on survey data and for users of survey data

Both, survey conditioning and panel conditioning, are the result of repeated measurement with the same or a similar instrument. The resulting changes in behavior and/or reporting, however, are unintended: In panel surveys, researchers usually assume that any intra-individual change over time reflects a 'true' change in respondents' behavior and/or reporting. That is, researchers assume that these changes would have also occurred had the respondent not participated in the survey (Warren and Halpern-Manners 2012). If some changes over time are caused by panel survey participation alone, however, they would not have occurred, had the respondent not participated in the survey. As a result, a sample will be less representative of the total population if panel conditioning results in behavioral changes. That is, after a few waves, the behavior of the sample interviewed does no longer represent the behavior of those people who were not interviewed and any statements about the respondents' behavior may be biased or wrong.

In addition, and this holds also for changes-in-reporting panel conditioning, researchers would over- or underestimate change over time because some of the change solely comes from the fact that respondents report differently in early waves of the panel than in later waves of the same panel. Consider, for example, the case from above, where researchers are interested in unemployed respondents' job search behavior and their employment probabilities. Interviewing respondents' about certain job search strategies may stimulate some respondents to engage in those job search strategies. At the same time, however, asking questions about unemployment and job search encourages respondents to spend more time on searching a new job, possibly resulting in respondents finding a new job faster than those who are not interviewed. Researchers might then conclude that changing job search strategies caused the decrease in the time needed to find a new job. In reality, however, the decrease in the time needed to find a new job is only caused by the stimulation resulting from being interviewed about their unemployment. Thus, researchers would underestimate the time that unemployed people need to find a new job and would (falsely) attribute the cause of the decrease in job search time to the change in search strategies.

Similarly, if survey conditioning results in respondents becoming better or worse reporters over time, then responses to questions asked early in the survey will be less (or more) accurate than answers to questions asked later in the survey. As a result, estimates derived from the data may under- or overestimate the underlying behavior.

Consider again the second example from above, where respondents misreport to eligibility questions to skip follow-up questions and thereby shorten the duration of the interview. If these questions refer to purchasing behavior, for example, then researchers would underestimate this behavior because some respondents underreport behavior in answers to questions asked later in the survey. As a result, any univariate estimates regarding purchasing behavior would be biased and would not well represent the true purchasing behavior of the whole population. In addition, if purchasing behavior was included in a multivariate model, e.g., to test a theoretical model, then any multivariate estimates derived from the model, such as regression coefficients, may be biased. In the worst case, researchers may falsely reject a model or falsely interpret empirical findings as evidence in favor of the model.

To sum up, once a respondent is conditioned by participating in several waves of a longitudinal survey or by responding to a set of similar questions in a single survey, subsequent reporting and/or behavior are/is a function of respondents' previous experience with (similar elements of) the same survey. As a result, estimates derived from the data may be biased and conclusions drawn from the data may be wrong.

1.4 Why this dissertation?

Concerns regarding the existence of conditioning effects have been documented in the literature since 1940 (e.g., Lazarsfeld 1940) and many scholars have studied survey and panel conditioning since then.[4] Considerable progress has been made regarding the circumstances *when* survey (e.g., Eckman et al. 2014) and panel conditioning (e.g., Waterton and Lievesley 1989; Warren and Halpern-Manners 2012) are likely to occur. Much less attention, however, has been paid to the question *how* to uncover such effects, especially those that occur in panel surveys. For instance, there are only a few studies that carefully distinguish between the different forms of panel conditioning (changes-in-reporting and changes-in-behavior). In fact, some studies claim to study changes-in-behavior panel conditioning, but they rely on respondents' self-reports only (e.g., Axinn, Jennings, and Couper 2015). Thus, they cannot tell whether survey participation actually affects respondents' behavior or if respondents simply report that their behavior changed. Furthermore, several other studies do not consider the possibility that other processes may lead to change in a panel survey over time, too (see, e.g., Das, Toepoel, and van Soest 2011, for a discussion). That is, these studies claim to identify a panel conditioning effect, but this effect may in fact be caused by another reason and not by survey participation.

The second chapter of this dissertation fills this gap and provides a methodological framework for the study and identification of the various forms of panel conditioning effects. Building on the methodological recommendations proposed in Chapter 2, I address the issue of disentangling changes-in-behavior from changes-in-reporting as well as the challenge of adjusting for other effects that may occur in panel surveys in Chapter 3. I am aware of only one other study (Crossley et al. 2017) that applies a methodologically sound research design to study changes-in-behavior due to survey participation. Thus, the importance of Chapter 3 lies in the contribution it makes to the more than sparse literature on changes-in-behavior panel conditioning.

Chapter 4 connects panel conditioning with survey conditioning. Although several studies have analyzed either one *or* the other form of conditioning, no study has analyzed the two in a joint

[4] See, e.g., Tourangeau, Kreuter, and Eckman (2015) for a literature review on survey conditioning effects and Warren and Halpern-Manners (2012) for a review of studies on panel conditioning effects.

context. But as some varieties of the two forms share the same theoretical mechanism and it is therefore evident to study the two together. Chapter 4, then, closes this gap by analyzing survey and panel conditioning in a single experiment.

The last of the three empirical chapters (Chapter 5) focuses on survey conditioning and answers the question whether survey conditioning depends on respondents' likelihood to participate in a survey. Only few studies have explored connections between respondents' likelihood to respond and other forms of misreporting (e.g., Olson 2006; Fricker and Tourangeau 2010), but the literature falls short of testing the connection with survey conditioning.

In sum, these four chapters make necessary contributions to various aspects of the literature on survey and panel conditioning, extending the current state of research in conditioning effects. The next - and last - section summarizes the contents of these chapters in more detail.

1.5 Summary of papers

Chapter 2 ("Methodological Challenges for the Analysis of Panel Conditioning Effects") provides an overview of three challenges confronting methodological research on panel conditioning and also of research designs developed to tackle them. First, as mentioned above, panel conditioning can lead to changes-in-reporting and changes-in-behavior. In some cases, both forms can even occur at the same time.[5] The challenge is that researchers working with self-reported behavior from survey data can hardly tell whether respondents' actual behavior changes over time or if it is only the reporting of behavior that changes. Thus, the first challenge I identify in Chapter 2 is to disentangle the two forms of panel conditioning. Second, in order to make statements about the causal effect of panel survey participation, researchers need to come up with control groups of people who have not been interviewed or have been interviewed only once. Finding such control groups, however, is often difficult and, hence, a second challenge. Third, other sources of error in (panel) surveys that result in changes in survey outcomes over

[5] For instance, a respondent may bring her behavior in line with society's norms after participating in several waves of a panel survey. At the same time, however, she might misreport this behavior due to social desirability in early waves of the panel, but report more honestly in later waves due to increased trust towards the interviewer.

time can easily be mistaken for panel conditioning. Panel attrition, for example, affects the composition of the respondent group because some respondents decide to not participate in all waves of a panel survey, resulting in changes in survey outcomes over time. These changes in survey outcomes over time may be misinterpreted as panel conditioning if they are not properly accounted for. Thus, the third challenge is to account for confounding sources of error when analyzing panel conditioning effects.

In addition to identifying these three challenges, I review and discuss research designs and methods developed to tackle them. To disentangle changes-in-reporting from changes-in-behavior, researchers can observe respondents' behavior in data that is independent of respondents' reporting. Control group data can easily be found in panel surveys that apply a rotating design, for example. The third challenge can be tackled by studying panel conditioning only among those respondents who respond to all waves of a panel or by conditioning on determinants of attrition, for example. This chapter concludes with a discussion of a future research agenda regarding panel conditioning effects.

In Chapter 3, I address all three challenges and study changes-in-behavior panel conditioning in the German panel study "Labor Market and Social Security" (Trappmann et al. 2013).[6] This panel study surveys recipients of unemployment benefits (amongst other groups) in Germany on an annual basis and has been running for more than ten years. One unique feature of this survey is that its respondents can be linked to administrative labor market records covering, inter alia, their phases of (un)employment, participation in labor market programs or unemployment benefits receipt. These records are independent of respondents' reporting; the panel survey (in combination with the administrative records) therefore offers an ideal setting to disentangle changes-in-behavior from changes-in-reporting. Starting out from the hypothesis that repeatedly answering questions about participation in active labor market policies (ALMP, i.e., programs designed to help unemployed people find a new job) increases respondents' take-up of these ALMP, I find that responding to several waves of the survey changes respondents' labor market behavior. That is, there is strong evidence that respondents' take-up of these programs

[6] The chapter is based on a joint paper with Stephanie Eckman. A revision of the paper is currently under review at the *Journal of the Royal Statistical Society, Series A: Statistics in Society* and an earlier version is published in the discussion paper series of the Institute for Employment Research (Bach and Eckman 2017).

increases as a result of repeated participation in the survey. To account for confounding sources of error (e.g., initial nonresponse, panel attrition and mode effects), I use an instrumental variable approach. I do so by selecting a control group of unemployment benefit recipients from the same administrative data. These recipients were also eligible for participation in the survey, but were never selected. Instrumenting actual participation in the survey among those selected for the survey with the random offer to participate in the survey allows us to adjust for confounding sources of error and identify a changes-in-behavior panel conditioning effect.

Chapter 4, by contrast, focuses on two other forms of conditioning effects.[7] In this chapter, I provide results regarding the question whether respondents of two consecutive waves of the Dutch Longitudinal Internet Study for the Social Sciences (Scherpenzeel 2011) misreport to filter questions, a form of survey conditioning (compare example in Section 1.2). Furthermore, I study whether this misreporting transfers to a second wave, leading to more misreporting in the second wave, which is a form of changes-in-reporting panel conditioning. Proceeding on previous research on misreporting to filter questions (e.g., Eckman et al. 2014), I hypothesize that respondents learn, after answering a few filter questions, that for each filter question triggered, several follow-up questions have to be answered. The motive behind this form of survey conditioning is the desire of some respondents to reduce the burden of the survey by answering fewer questions. Similarly, I hypothesize that this learning effect transfers to a second wave fielded one month after, resulting in increased conditioning effects in the second wave. Results from this study suggest that respondents learn to misreport in a single wave, but this learning does not transfer to a second wave. To sum up, there is evidence for survey conditioning, but there is no evidence for changes-in-reporting panel conditioning.

Chapter 5 builds on and extends parts of the analysis of Chapter 4. I replicate experimental results regarding misreports to two forms of eligibility questions due to survey conditioning in five surveys conducted in the Netherlands, the U.S. and Germany (Eckman et al. 2014; Bach and Eckman forthcoming; Eckman and Kreuter forthcoming)[8]. The core contribution of this chap-

[7] The chapter is based on a joint paper with Stephanie Eckman. The paper has been accepted for publication in the *Journal of Survey Statistics and Methodology* (Bach and Eckman forthcoming).

[8] The chapter is based on a joint paper with Stephanie Eckman and Jessica Daikeler. The paper is under review in *Public Opinion Quarterly.*

ter, however, is the extension of their analyses to the question whether *reluctant* respondents are worse reporters than likely respondents, i.e., whether reluctant reporters are more susceptible to becoming worse reporters due to survey conditioning. The nonresponse-measurement error model developed by Groves (2006) predicts a nexus between response propensity (the probability to respond to a specific survey) and measurement error (e.g., misreporting due to survey conditioning). There may be some common causes that influence both the decision to participate in a survey and the degree of misreporting. In this study, I hypothesize that the desire to reduce the burden of the survey may be such a common cause. That is, respondents who have a strong desire to reduce the burden of the survey have both a low probability to respond to the survey in the first place and a high desire to misreport during the interview to reduce the duration of the survey. I identify reluctant respondents by estimating respondents' response propensities in each of the five surveys. Those respondents with the lowest response propensities in each survey are reluctant respondents. Comparing misreporting due to survey conditioning between the most reluctant and the most likely respondents, I do not find consistent evidence that response propensity influences misreporting due to survey conditioning. Only in one survey do I find reluctant respondents to be worse reporters. Thus, our results suggest that reluctant respondents do not report worse to eligibility questions due to survey conditioning.

References

Axinn, William G., Elyse A. Jennings, and Mick P. Couper (2015). "Response of Sensitive Behaviors to Frequent Measurement". In: *Social Science Research* 49, pp. 1–15.

Bach, Ruben L. and Stephanie Eckman (forthcoming). "Motivated Misreporting in Web Panels". In: *Journal of Survey Statistics and Methodology*.

— (2017). "Does Participating in a Panel Survey Change Respondents' Labor Market Behavior?" In: *IAB Discussion Paper* 15/2017. Available at `http://doku.iab.de/discussionpapers/2017/dp1517.pdf`, last accessed Jan 23, 2018.

Crossley, Thomas, Jochem de Bresser, Liam Delaney, and Joachim Winter (2017). "Can Survey Participation Alter Household Saving Behavior?" In: *Economic Journal* 127, pp. 2332–2357.

Das, Marcel, Vera Toepoel, and Arthur van Soest (2011). "Nonparametric Tests of Panel Conditioning and Attrition Bias in Panel Surveys". In: *Sociological Methods & Research* 40.1, pp. 32–56.

Duan, Naihua, Margarita Alegria, Glorisa Canino, Thomas McGuire, and David Takeuchi (2007). "Survey Conditioning in Self-Reported Mental Health Service Use: Randomized Comparison of Alternative Instrument Formats". In: *Health Research and Educational Trust* 42.2, pp. 890–907.

Eckman, Stephanie and Frauke Kreuter (forthcoming). "Misreporting to Looping Questions in Surveys: Recall, Motivation and Burden". In: *Survey Research Methods.*

Eckman, Stephanie, Frauke Kreuter, Antje Kirchner, Annette Jäckle, Roger Tourangeau, and Stanley Presser (2014). "Assessing the Mechanisms of Misreporting to Filter Questions in Surveys". In: *Public Opinion Quarterly* 78.3, pp. 721–733.

Fricker, Scott and Roger Tourangeau (2010). "Examining the Relationship between Nonresponse Propensity and Data Quality in Two National Household Surveys". In: *Public Opinion Quarterly* 74.5, pp. 934–955.

Groves, Robert M. (2006). "Nonresponse Rates and Nonresponse Bias in Household Surveys". In: *Public Opinion Quarterly* 70.5, pp. 646–675.

Lazarsfeld, Paul F. (1940). "'Panel Studies'". In: *The Public Opinion Quarterly* 4, pp. 122–128.

Olson, Kristen (2006). "Survey Participation, Nonresponse Bias, Measurement Error Bias, and Total Bias". In: *Public Opinion Quarterly* 70.5, pp. 737–758.

Scherpenzeel, Annette (2011). "Data Collection in a Probability-Based Internet Panel: How the LISS Panel Was Built and How It Can Be Used". In: *BMS: Bulletin of Sociological Methodology / Bulletin de Méthodologie Sociologique* 109, pp. 56–61.

Tourangeau, Roger, Frauke Kreuter, and Stephanie Eckman (2015). "Motivated Misreporting: Shaping Answers to Reduce Survey Burden". In: *Survey Measurements. Techniques, Data Quality and Sources of Error.* Ed. by U. Engel. Frankfurt/New York: Campus, pp. 24–41.

Trappmann, Mark, Jonas Beste, Arne Bethmann, and Gerrit Müller (2013). “The PASS Panel Survey After Six Waves”. In: *Journal for Labour Market Research* 46.4, pp. 275–281.

Warren, John R. and Andrew Halpern-Manners (2012). “Panel Conditioning in Longitudinal Social Science Surveys”. In: *Sociological Methods & Research* 41.4, pp. 491–534.

Waterton, Jennifer and Denise Lievesley (1989). “Evidence of Conditioning Effects in the British Social Attitudes Panel”. In: *Panel Surveys*. Ed. by D. Kasprzyk, G. J. Duncan, G. Kalton, and M. P. Singh. New York: Wiley, pp. 319–339.

Chapter 2

Methodological Challenges for the Analysis of Panel Conditioning Effects

2.1 Abstract

Panel conditioning refers to the phenomenon whereby respondents' attitudes, behavior, reporting of behavior and/or knowledge are changed by repeated participation in a panel survey. Uncovering such effects, however, is difficult due to three major methodological challenges. First, researchers need to disentangle changes in behavior from changes in the reporting of behavior as panel conditioning may result in both, even at the same time and in opposite directions. Second, the identification of the *causal* effect of panel participation on the various forms of change mentioned above is complicated as it requires comparisons of panel respondents with control groups of people who have not been interviewed before. Third, other sources of error in (panel) surveys may easily be mistaken for panel conditioning if not properly accounted for. Such error sources are panel attrition, mode and interviewer effects. We review the challenges mentioned above in detail and discuss the strengths and weaknesses of the various designs that researchers have developed to address them. We conclude with a discussion of a future research agenda on panel conditioning effects in longitudinal surveys.

2.2 Introduction

Researchers working with data collected through surveys usually assume that the act of measuring does not affect what is being measured. Yet, as early as in 1940, Lazarsfeld (1940, p.128) noted that "the big problem, as yet unsolved, is whether (...) interviews are likely, in themselves, to influence a respondent's opinion". Since then, researchers have spent decades analyzing how repeated measurement can result in unintended changes in respondents' attitudes, knowledge, behavior and reports of behavior. These changes are unintended insofar as researchers usually assume that any intra-individual over-time change found in their data reflects a 'true' change in a person's attitudes or behavior, and that these changes would have also occurred had the respondent not participated in the survey (Warren and Halpern-Manners 2012). Yet, when the sheer act of measuring affects what is being measured, this assumption is no longer satisfied and researchers risk mis-characterizing the existence, magnitude, and correlates of changes across survey waves in respondents' attitudes and behaviors (Clinton 2001; Halpern-Manners, Warren, and Torche 2014).

Although scholars from various disciplines have spent much time on uncovering these so-called *panel conditioning* effects, there are three methodological challenges that have plagued researchers of panel conditioning ever since. First, panel conditioning may influence both respondents' reporting of behavior and the actual behavior itself, even at the same time and in opposite directions. Therefore, researchers need to disentangle changes in respondents' behavior from changes in respondents' reporting of behavior to get unbiased estimates of either one. For example, respondents may remember from participation in prior waves of a panel survey how the interview is structured and how they can speed through the interview by taking shortcuts. We call this form of panel conditioning changes-in-reporting panel conditioning. At the same time, respondents' actual behavior may be conditioned by participation in the survey. For example, repeatedly answering questions may work as a stimulus that affects respondents' subsequent behavior. We call this form of panel conditioning changes-in-behavior panel conditioning.[1] To get unconfounded estimates of either form of panel conditioning, it is crucial to

[1] We will discuss examples of both types of panel conditioning in detail throughout this paper.

clearly distinguish between the two.

The second challenge refers to the availability of control group data. That is, researchers of panel conditioning (at least implicitly) wish to estimate the *causal effect* of panel survey participation on respondents' attitudes, knowledge, reporting of behavior or actual behavior (Bach and Eckman 2017). To estimate a causal or treatment effect, we usually compare treated cases against untreated, i.e., control cases in (quasi) experimental designs. In the panel conditioning framework, the former are those cases of individuals who responded to two or more waves of a panel survey, while the latter are cases who have not been interviewed or have been interviewed only once. To identify an unbiased treatment effect, assignment to treatment and control is, ideally, random because randomization balances (in expectation) all differences (e.g., sociodemographic differences) between the treatment and the control group. As a result, the only remaining difference between the two groups is that one receives the treatment, and the other one does not. In large-scale social science panel surveys, the kind of surveys that social scientists frequently use for substantive research, methodological experiments with random assignment of cases to treatment or control (i.e., participation in a survey several times or only once) are hardly ever (intentionally) implemented (Warren and Halpern-Manners 2012). That is, data for treated cases (panel survey respondents) are usually easily available from the panel survey itself; control group data, however, are much harder to find. As a result, the identification of the causal effect is difficult without further assumptions.

The third challenge is that, even when experimental manipulations can be implemented, researchers need to account for confounding sources of error. Error sources that may confound estimates of panel conditioning are longitudinal nonresponse (i.e., panel attrition), mode effects and interviewer effects. While attrition affects the composition of a longitudinal survey sample over time, mode effects and interviewer effects confound the level of measurement error. All of them may result in changes in an outcome variable over time that are not due to a 'true' change in a person's attitudes or behavior and may thus easily be mistaken for panel conditioning. As a result, estimates of panel conditioning, even when based on experimental manipulations, will be biased if these error sources are not properly accounted for.

In the remainder of this article, we briefly summarize the different forms of panel conditioning effects and hypotheses why panel conditioning can occur in social science longitudinal surveys. We then elaborate on the methodological challenges mentioned above and the research designs developed to tackle them. We do so by demonstrating for each challenge why it requires researchers' special attention when studying panel conditioning and by discussing the research designs and methods developed by previous research to deal with them. We conclude with a discussion of a future research agenda for panel conditioning effects in social science longitudinal surveys.

2.3 Various forms of panel conditioning

Panel conditioning has been studied for some decades. However, the literature often lacks a clear distinction between the various forms of panel conditioning. Moreover, the underlying causal mechanisms as well as their theoretical foundations are often unclear. To date, there is no unified theory explaining the nature and the mechanisms underlying panel conditioning. However, several authors have proposed hypotheses under which panel conditioning is likely to affect respondents' attitudes, knowledge, behavior and reporting of behavior (Waterton and Lievesley 1989; Cantor 2010; Warren and Halpern-Manners 2012). We distinguish between the different forms of panel conditioning below, review hypotheses that explain when the different forms may arise and present empirical evidence.

2.3.1 Changes-in-reporting panel conditioning

Changes-in-reporting panel conditioning refers to the phenomenon whereby panel survey participation influences the way respondents report over time. This form of conditioning can result in respondents becoming either better or worse reporters over the waves of a panel survey. Respondents can become better reporters because they become more trusting of the survey experience. For example, they might feel more comfortable and more motivated giving less socially desirable, but more accurate answers. That is, increasing trust in the confidentiality of their responses and reduced suspicion towards the interviewer leads to a reduction in measurement error over time as respondents report more honestly. Similarly, respondents may gain a better understanding of the meaning of the questions or may become more convinced of the importance of their answers for the survey and report fewer "don't know"-responses. Thus, respondents become more comfortable with the interviewing process leading them to answer the questions more accurately (Bailar 1989; Waterton and Lievesley 1989; Van der Zouwen and Van Tilburg 2001; Warren and Halpern-Manners 2012).

The literature documents several examples of changes-in-reporting panel conditioning resulting in respondents becoming better reporters. Halpern-Manners, Warren, and Torche (2014) find that respondents report more honestly regarding having previously driven drunk or having stolen something of little value in later waves of a panel survey. Similarly, Waterton and Lievesley (1989) report declines in social desirability bias (reporting racial prejudice) over time and fewer reports of "don't knows". Chadi (2013) finds that increased trust towards the interviewer leads to more honest reporting of life satisfaction, a finding replicated by Van Landeghem (2012). Furthermore, Kroh, Winter, and Schupp (2016) find that the reliability of person fit measures increases with every wave of a panel survey because respondents get a better understanding of the interview process. Angel, Heuberger, and Lamei (2017) report that both over- and under-reporting of household income decreases over panel waves as respondents feel more comfortable reporting their income honestly and prepare for the interviews to provide more accurate responses to the survey.

On the other hand, respondents may become worse reporters over the waves of a panel survey (Bailar 1989; Waterton and Lievesley 1989; Williams, Block, and Fitzsimons 2006; Warren and Halpern-Manners 2012). Respondents can learn from participation in previous waves of a survey how the interview is structured and how its burden or length can be reduced by taking shortcuts in the interview. For example, respondents of a labor market survey might learn that reporting to be employed leads to additional follow-up questions regarding the employment and may report to be unemployed in later waves to skip the follow-up questions. As a result, measurement error will increase over the waves of a panel survey. Declining data quality over time may also result from increasing levels in social desirability bias. That is, when questions refer to socially non-normative behavior, respondents may be confronted with a conflict between their behavior and society's norms. To avoid such cognitive dissonance, respondents may resort to reporting behavior closer to society's norms.

Examples of respondents becoming worse reporters due to panel conditioning are numerous. Schonlau and Toepoel (2015), for example, show that straightlining, i.e., the tendency to give the same responses to a series of questions with identical answer choices, increases with respondents' panel experience. Several other studies find increases in social desirability bias over time, e.g., in reports of children's vaccination status (Battaglia, Zell, and Ching 1996) or in reports of exercising (Williams, Block, and Fitzsimons 2006). Furthermore, it is well documented that the Current Population Survey (CPS) underestimates unemployment rates because respondents misreport unemployment to skip follow-up questions in later waves of the survey (Hansen et al. 1955; Bailar 1975; Bailar 1989; Shack-Marquez 1986; Solon 1986; Shockey 1988; Halpern-Manners and Warren 2012).[2] Other studies find respondents taking shortcuts in later waves of a panel survey regarding reports of home alteration and repair jobs (Neter and Waksberg 1964), functional limitations among elderly people (Mathiowetz and Lair 1994), substance abuse (Torche, Warren, and Halpern-Manners 2012), every day personal hygiene product use (Nancarrow and Cartwright 2007) and consumption of purchased and own-produced goods (Schündeln 2017). Similar effects, called *survey conditioning*, may also occur in cross-sectional surveys. For example, respondents learn to misreport to filter questions in order to skip follow-up questions and speed through an interview (e.g., Duan et al. 2007; Kreuter et al. 2011; Eckman et al. 2014; Eckman and Kreuter forthcoming; Bach and Eckman forthcoming). However, there are also studies that find no increases or decreases in misreporting over time in panel surveys (e.g., Cohen and Burt 1985; Halpern-Manners, Warren, and Torche 2014; Struminskaya 2016; Bach and Eckman forthcoming).

[2] Some of these studies likely over- or underestimate the panel conditioning effect in the CPS due to confounding panel attrition and mode effects (see Section2.4.3). However, the general finding that the CPS underestimates unemployment rates seems to hold even when attrition and mode effects are accounted for (Halpern-Manners and Warren 2012).

2.3.2 Changes-in-behavior panel conditioning

Panel participation may also result in changes in actual behavior over time. The common explanation for this type of panel conditioning is the cognitive stimulus approach. It holds that repeatedly being asked the same questions causes respondents to become more aware of the topic of the survey, raises their consciousness of the issues and motivates them to engage in the behavior under study (Sturgis, Allum, and Brunton-Smith 2009; Waterton and Lievesley 1989; Warren and Halpern-Manners 2012). For example, the participation in a pre-election poll may increase voter turnout because being asked about voting intentions increases the likelihood of actual voting.

Furthermore, being asked knowledge questions might stimulate respondents who do not know the answer to look it up after the interview. E.g., respondents of a political science survey who are not aware of a new law might look it up after the survey. In a follow-up interview, they would then know the correct answer. Their knowledge would not have changed, however, had they not participated in the interview. Similarly, survey questions can serve to provide information about behaviors that respondents were not aware of. If respondents had not participated in a survey, they would not have known of the behavior under study and could not have engaged in it (Halpern-Manners and Warren 2012). A survey that asks for participation in a specific cancer screening measure, for example, might inform a respondent who had not known of the measure before about the existence of this measure. Therefore, had the respondent not participated in the survey, she would not have known of that measure.

Changes in behavior may also arise from survey questions dealing with socially non-normative or stigmatized behavior or attitudes. As respondents are confronted with a conflict between their attitudes or behavior and society's norms, they bring their future behavior or attitudes in line with society's norms (Williams, Block, and Fitzsimons 2006).[3] In a survey about alcohol consumption, for example, dissonance between respondents' heavy-drinking habits and society's norms of modest alcohol consumption may stimulate respondents to reconsider the amount and

[3] Note that respondents of surveys containing questions asking for socially non-normative behavior may be subject to both changes-in-reporting as well as changes-in-behavior panel conditioning. This dual effect poses a major challenge to panel conditioning research (see Sections 2.2 and 2.4.1).

frequency of alcohol consumption and thereby lead them to drink less by the time of a follow-up survey (Warren and Halpern-Manners 2012).

Panel participation might also lead to 'real' changes in an attitude. When attitudes are less crystallized, responding to questions about an attitude can sometimes change it. Moreover, when respondents lack crystallized attitudes about a specific topic, they will nonetheless offer a response to a question about that attitude. This response may start a cognitive process that will lead to a change in the attitude by the time of the next wave (Waterton and Lievesley 1989; Sturgis, Allum, and Brunton-Smith 2009; Warren and Halpern-Manners 2012). For example, respondents of a survey on same-sex marriage may not have an opinion regarding this topic, but might provide an opinion nonetheless. Moreover, responding to a question regarding their views on this topic may also start a cognitive process that will lead them to form an opinion.

Changes-in-behavior panel conditioning is reported in studies of voting behavior, where participation in a pre-election survey leads to increases in voter turnout in upcoming elections (Clausen 1969; Kraut and McConahay 1973; Yalch 1976; Traugott and Katosh 1979; Granberg and Holmberg 1992). However, not all studies detect this effect (Smith, Gerber, and Orlich 2003). Other behaviors affected by panel conditioning are water treatment product use (Zwane et al. 2011); purchases of health insurance (Zwane et al. 2011), automotive services (Borle et al. 2007), automobiles (Morwitz, Johnson, and Schmittlein 1993; Chandon, Morwitz, and Reinartz 2005) and computers (Morwitz, Johnson, and Schmittlein 1993); saving for retirement (Crossley et al. 2017); cheating in exams (Spangenberg and Obermiller 1996), use of contraceptives (Axinn, Jennings, and Couper 2015); perceptions of marital quality (Veroff, Hatchett, and Douvan 1992) and participation in labor market programs (Bach and Eckman 2017). Regarding panel conditioning leading to changes in knowledge, increases in knowledge are reported for bacteria and for pension schemes (Das, Toepoel, and van Soest 2007), contraception methods (Coombs 1973) and vaccination programs (Battaglia, Zell, and Ching 1996).

To sum up, theory and evidence from numerous studies suggest that repeated participation in panel surveys can lead to changes in respondents' reports of behavior and attitudes as well as to changes in their actual behavior, attitudes and knowledge. Furthermore, all forms of panel conditioning are more likely to occur the shorter the interval between the waves of a panel study is and the more often respondents are interviewed (Halpern-Manners, Warren, and Torche 2014; Van Landeghem 2012). Identifying a panel conditioning effect of any form, however, is difficult, as we will see in the next section.

2.4 Methodological challenges

Three major methodological challenges (disentangling the different forms of panel conditioning, finding control group data and accounting for confounding sources of error) complicate the identification of panel conditioning effects in longitudinal data. We review them in detail in this section. Moreover, we discuss the various designs that previous studies have developed to address them. Table 2.1 summarizes the methodological challenges and the approaches developed to tackle them.

2.4.1 Challenge 1: Disentangling the different forms of panel conditioning

Repeated participation in a panel survey can result in both changes-in-reporting and changes-in-behavior as we have laid out above.[4,5] Surveys, however, usually measure behavior through respondents' self-reports only. Thus, working with survey data only, i.e., with respondents' self-reported behavior, it is difficult for researchers to tell whether panel conditioning affects respondents' behavior or reports of behavior. To further complicate things, panel conditioning may even affect both at the same time. Bach and Eckman (2017), for example, discuss a hypothetical example where respondents of a panel survey on recycling and environmental behavior over-report recycling in early waves due to social desirability, but report more honestly

[4] We do not need to differentiate between different effects of panel survey participation on respondents' knowledge. Panel conditioning either increases respondents' knowledge or does not affect it at all.

[5] Strictly speaking, one would need to disentangle changes in the reporting of attitudes from real changes in attitudes in a similar way because panel conditioning can result in both (real changes in an attitude and changes in the reporting of an attitude (see previous section). Since attitudes are, to the author's knowledge, always measured through self-reports, disentangling the two is impossible.

Table 2.1: Summary of methodological challenges and solutions

Challenge	**Solutions**	**Remarks**
Challenge 1: Disentangling changes-in-behavior and changes-in-reporting	Records independent of respondents' reporting (e.g., administrative data) to study changes-in-behavior	Requires data independent of respondents' reporting and linkage with respondents
	Validation records independent of respondents' reporting (e.g., administrative data) to study changes-in-reporting	Requires data independent of respondents' reporting and linkage with survey records
	Exclude changes-in-behavior based on theoretical considerations	
Challenge 2: Finding control group data	*Within-person approach*: Use panel respondents as their own control group	Restricted to analysis of some forms of changes-in-reporting (e.g., measurement error)
	Between-person approaches	
	Variant 1: Implement experiment	
	Variant 2: Placebo interviews	Confounding effects of placebo interviews?
	Variant 3: Cross-sectional survey with same content	Cross-sectional survey: same mode, same questions, same sampling design and population as well as same time of data collection
	Variant 4: Panel survey with rotating designs	
Challenge 3: Accounting for confounding sources of error		
Panel attrition	Disregard respondents who do not participate in all waves	Only in combination with within-person approach and variants 2 and 4 of between-person approach
	Condition on observable determinants of attrition	Requires (untestable) missing-at-random assumption
	Instrumental variables	Requires outcome data for both respondents and nonrespondents as well as control group of people not interviewed
Mode and interviewer effects	Restrict to cases interviewed in same mode and by same interviewers	

in later waves as they become more comfortable and trusting of the interview process. At the same time, however, repeatedly asking respondents about recycling and environmental behavior may work as a stimulus and increase respondents' awareness of the importance of recycling and thereby lead to changes in their behavior. Researchers seeking to understand panel conditioning need to be aware of the various ways that panel conditioning might affect behavior and the reporting of it at the same time. It is therefore crucial to clearly distinguish between the two forms (Waterton and Lievesley 1989; Van der Zouwen and Van Tilburg 2001).

Luckily, disentangling changes-in-reporting from changes-in-behavior is straightforward when records that are unaffected by respondents' reporting are available, for example, from administrative records. Using such data, researchers can study changes-in-behavior by linking both data sources and analyzing respondents' behavior in the administrative records only. Such administrative data (e.g., tax records or insurance records), are themselves not free of error (e.g., Oberski et al. forthcoming). Yet, they can be considered the gold standard for analyzing changes-in-behavior panel conditioning because they are usually generated independently of respondents' reporting.

Linked survey-administrative records can also be used to study changes-in-reporting, for example, using them to validate survey responses. If measurement error, the deviation of a survey report from the true value (i.e., the value recorded in the validation data), changes over time in a panel survey, we can interpret such results as evidence for changes-in-reporting panel conditioning.

Unfortunately, however, in most scenarios, researchers will not have external validation records at hand, thereby making disentanglement of the two forms of panel conditioning difficult or impossible. In a few cases, theoretical considerations may allow researchers to conclude that only one of the two forms of panel conditioning is possible. Warren and Halpern-Manners (2012), for example, conclude that respondents of the CPS who report to be unemployed in the first wave, but report to be out of the labor force in subsequent waves, do so to avoid follow-up questions and *not* because participating in the survey made them actually more likely to leave the labor force. Thus, some scenarios may allow researchers to preclude one or the other form of panel conditioning.[6]

It is likely that the lack of administrative records explains why many studies do not consider differences between changes-in-reporting and changes-in-behavior panel conditioning. If researchers do not clearly disentangle the two forms, they risk making flawed statements about either form. In the worst case, the two forms cancel each other out, leading researchers to conclude that panel conditioning is not present.

Only a few studies explicitly acknowledge that panel survey participation can lead to both forms of change. Angel, Heuberger, and Lamei (2017), studying the development of misreports of household income over time (see Section 2.3.1), for example, link respondents' survey reports of their household income to register data containing the same information. Using the income from the register as validation records, they are able to study how panel survey participation leads to changes-in-reporting. Similarly, Yan and Eckman (2012) link survey responses of a large-scale labor market panel survey to administrative labor market records to disentangle the two forms of panel conditioning, demonstrating that both take place at the same time. Crossley et al. (2017) and Bach and Eckman (2017) are two examples of studies that explicitly study changes-in-behavior panel conditioning by observing the behavior of respondents of large-scale social science panel surveys in administrative records. Similar approaches are applied by the other studies cited in Section 2.3.2, though many of them rely on small datasets collected among students (e.g., Spangenberg and Obermiller 1996) or marketing data (e.g., Chandon, Morwitz, and Reinartz 2005) and do not use the kind of longitudinal data that builds the basis for substantive research in the social sciences.

To sum up, disentangling the two forms of panel conditioning is difficult when administrative data or validation records of reported behavior do not exist or cannot be linked. In these cases, no clear statement can be made about either form of panel conditioning. Recent studies (e.g., Crossley et al. 2017; Angel, Heuberger, and Lamei 2017), however, have made considerable methodological advances in disentangling the different forms of panel conditioning.

[6] While excluding the possibility that survey participation affects actual behavior seems justified in some scenarios, we cannot think of an example where panel conditioning may result in changes-in-behavior only, i.e., excluding changes-in-reporting.

2.4.2 Challenge 2: Finding control groups

The second challenge (finding control group data) can be best understood in terms of the counterfactual causal model (e.g., Holland 1986). Researchers analyzing panel conditioning usually wish to study the *causal* effect of panel survey participation. Thus, we can think of panel conditioning as the effect of a treatment, $D \in [0, 1])$ (panel survey participation) on some outcome, Y (e.g., the value of a reported survey variable). To define the treatment effect, τ, we define two potential outcomes, following Rubin (1974) and Rubin (1978): $Y_{i,1}$ is the outcome that occurs when a case i receives treatment (participates in the survey), $Y_{i,0}$, by contrast, is the outcome when a case does not receive treatment, i.e., is in the control condition (does not participate in the survey). Using these potential outcomes, we define the individual treatment effect as $\tau_i = Y_{i,1} - Y_{i,0}$. The *fundamental problem of causal inference*, however, is that we observe only $Y_i = D_i Y_{i,1} + (1 - D_i) Y_{i,0}$ for any individual (Holland 1986). In other words, we observe only one of the two potential outcomes because a person either participates in a survey or does not. While this problem prevents us from estimating an individual treatment effect, we may still estimate an average treatment effect, such as the average treatment effect on the treated (ATT), $\tau_{ATT} = E(\tau|D = 1) = E(Y_1|D = 1) - E(Y_0|D = 1)$.

In the panel conditioning case, the expected value of the outcome of the treated cases, $E(Y_1|D = 1)$, can be directly observed from the data: it is simply the expected value of Y among respondents who responded to several waves of a panel. The counterfactual outcome of the treated, $E(Y_0|D = 1)$, by contrast, cannot be observed. When treatment assignment is random, however, we can replace $E(Y_0|D = 1)$ with $E(Y_0|D = 0)$, the expected value of the outcome among the cases that were not treated (that is, the expected value of Y of the control group), and estimate an unbiased treatment effect. Using this framework to study panel conditioning effects, the challenge is to find suitable estimates of the control group outcome, i.e., an estimate of Y from persons who did not participate or participated only once in a survey, but who would participate in all waves of the panel, had they been selected.

Approaches to replace the counterfactual outcome of the treated can broadly be grouped into two categories. The first approach relies on within-person comparisons and the second on between-person comparisons. A review of the literature suggests that the majority of studies applies a variant of the between-person approach. Each approach has its own specific strengths and weaknesses in identifying panel conditioning effects. One challenge that all of them face is eliminating or adjusting for panel attrition (one aspect of the third methodological challenge, see Section 2.4.3). Our initial review of the different ways to come up with an estimate of the counterfactual does not include a discussion of panel attrition. Instead, we discuss panel attrition and other sources of confounding error only after all approaches haven been reviewed.

The within-person approach

The within-person approach uses panel respondents as their own control group and simply compares survey outcomes among the same people at different waves of a panel. That is, the control group outcome is usually defined as responses in the first wave of a panel (where respondents are not yet conditioned), while the outcome of the treatment group is defined as responses by the same respondents to subsequent waves of the same panel survey. A few studies have applied this design, e.g., by comparing coefficients of variation at different waves of a panel diary (Toh, Lee, and Hu 2006), by comparing responses from a baseline survey with follow-up reports of the same respondents collected one week later (Sharpe and Gilbert 1998), by studying trends in attitudes across eleven waves of the British Household Panel Survey (Sturgis, Allum, and Brunton-Smith 2009) or by analyzing person fit measures in several dozen waves of the German Socio-Economic Panel (Kroh, Winter, and Schupp 2016).

The ability of this method to estimate an unbiased panel conditioning effect, however, heavily depends on the object of study. If one is interested in behavioral or attitudinal changes due to panel conditioning, one needs to account for the fact that there are many other factors (besides participating in the panel survey) that may cause a change in behavior or attitudes. Disentangling such 'true' change (i.e., change, that would have happened even in the absence of participation in the survey) from change that is caused by survey participation alone is impossible when relying on a single set of respondents only (see, for example, Shadish, Cook, and Campbell 2002, ch.4, for a general discussion of this approach in the causal inference framework). Thus, this method will lead to biased estimates of panel conditioning in many settings. Sturgis, Allum, and Brunton-Smith (2009), for example, conclude from discovering changes over time in political attitudes among the same respondents of the British Household Panel Survey that these attitudes changed due to panel conditioning. However, many factors that influence political attitudes changed over the same course of time (e.g., new bills being passed or new governments being elected). Thus, change in political attitudes over several waves of the panel is not only caused by panel survey participation, but also by change in other (external) factors over time. Disentangling these two forms of change, however, is impossible using only one set of respondents. That is, Sturgis, Allum, and Brunton-Smith's (2009) estimates of panel conditioning are likely biased because they do not separate change in attitudes over time that would have happened irrespective of panel participation from change over time that is only due to panel participation.

In other scenarios, relying on a single set of respondents can work well, however. Angel, Heuberger, and Lamei (2017), for example, study changes in the level of measurement error in income reports over time. Other than attitudes or behavior, which can change due to external influences (see above), measurement error in respondents' survey reports is unlikely to covary with external factors. That is, if the level of measurement error is smaller (or larger) in the first wave than in subsequent waves, respondents became worse (or better) reporters over time due to panel conditioning. Thus, when the objective of a study is to analyze changes-in-reporting panel conditioning, relying on a single set of respondents can produce unbiased estimates of panel conditioning effects.

The between-person approach

The second method taking a different approach is more powerful in this regard because it can be applied to study *all* forms of panel conditioning. Instead of within-person comparisons, it relies on between-person comparisons, usually implemented in an experimental design with random assignment of cases to either the treatment or the control group (see the potential outcomes framework introduced above). That is, the between-person approach compares outcomes between respondents with varying levels of exposure to the treatment of survey participation. There are many different variants of this design, which we review below.

The first variant randomly assigns people to either repeated participation in a panel survey or to one-time participation in one wave of the same survey only. As a result, some respondents are interviewed several times (treatment condition), whereas other respondents are interviewed only once (control condition). Kruse et al. (2009), for example, compare attitudes and (reported) behavior of respondents from eleven subsequent waves of the Knowledge Networks Panel with attitudes and behavior of respondents of three independent cross-sectional samples of the same panel that were interviewed with the same instrument at waves three, six or nine. Using a similar design, Axinn, Jennings, and Couper (2015) compare reported contraception use between women who were assigned to a baseline survey followed by weekly journal keeping for twelve months and women who were assigned to the baseline survey and a follow-up survey after twelve months only. Pushing the variant described in this paragraph to the most extreme, three studies (Zwane et al. 2011; Crossley et al. 2017; Bach and Eckman 2017) randomly assign people to participation in (several waves of) a survey (treatment) or to no participation at all (control). To measure the outcome, they identify survey participants and people not assigned to survey participation in administrative records that contain measures of the outcome and are available for both groups. In the absence of nonresponse and panel attrition, any differences between the sample selected for survey participation and the control group sample that is observed in the administrative records only are then due to panel conditioning.

The second variant is similar to the first, with one important difference. In variant one, the treatment group is interviewed several times with the same instrument, whereas the control group is interviewed only once (or not at all). In the second variant, however, both the treatment and the control group are interviewed several times, but with different instruments. For example, a study that analyzes the (panel conditioning) effect of repeatedly answering a set of questions A may split respondents into two groups. One group is interviewed with questions A for several waves and the other group is interviewed with questions B. Only in the last wave is the latter group also interviewed with questions A. As a result, the first group is treated with A, but the second one not. Such a design is implemented, for example, by Struminskaya

(2016). The main reason for interviewing the control group with other or 'placebo' questions instead of not interviewing them at all is that many panel surveys simply cannot afford to *not* interview some respondents.

The third variant of the between-respondent approach builds on a somewhat different idea. Instead of assigning people to different levels of exposure to the same survey (or variants of it), this approach uses data from two different surveys, where one is a panel survey and the other is a cross-sectional survey. Wilson and Howell (2005), for example, compare trends in the prevalence of arthritis in the U.S. between 1992 and 2002 derived from panel respondents of the Health and Retirement Survey (HRS) with arthritis trends for the same years derived from cross-sectional respondents of the National Health Interview Survey (NHIS). Because data from the two surveys are comparable, as the authors argue, any differences between trends in the HRS and the NHIS data (in the absence of panel attrition in the HRS) are due to panel conditioning in the former. In other words, because respondents of the HRS respond to the same survey multiple times (the treatment group) and the NHIS is a repeated cross-sectional survey that interviews a new group of respondents in each round (the control groups), any differences in the prevalence of arthritis trends between the two is due to panel conditioning in the HRS.

The fourth variant exploits experimental manipulations that occur unintentionally in some panel surveys. Because the sample size of panel surveys decreases over time due to panel attrition, some surveys introduce refreshment samples from time to time to renew the respondent pool to approximately its original sample size (e.g., the General Social Survey or the Current Population Survey in the U.S. or the panel study Labor Market and Social Security in Germany). To ensure comparability between the original sample and the refreshment sample(s), refreshment samples are usually drawn from the same population and with the same sampling design. Thus, people who join a panel survey as part of a refreshment sample form an ideal control group: When they join the panel as novice respondents, respondents of the original sample have already participated in several waves of the panel. Thus, the only difference between the two groups (absent of panel attrition in the original sample) is that one group has been treated and the other one has not been treated (yet). Warren and Halpern-Manners (2012), for example, use this design to study panel conditioning effects in several outcomes measured in the General Social Survey and the German Socio-Economic Panel.

A slightly different variant of this approach has been applied extensively to study panel conditioning effects among respondents of the Current Population Survey (Hansen et al. 1955; Bailar 1975; Shack-Marquez 1986; Solon 1986; Shockey 1988; Bailar 1989; Halpern-Manners and Warren 2012). This survey uses a rotating panel structure, that is, respondents are only interviewed for a certain number of waves before being removed from the panel. At each wave, a new rotation group joins the panel. Comparing outcomes between respondents of one rotation group with another rotation group (with different panel tenure) thus offers an ideal setting to study panel conditioning effects.

All of the four variants reviewed above can produce unbiased estimates of the counterfactual outcome $E(Y_0|D = 1)$ (had the respondents not participated in the survey). Variants one and four seem the most powerful as they do not require as many assumptions as variants two and three. At the same time, however, they are also the most difficult to implement. Variant one requires researchers to intentionally implement experimental manipulations in a survey. Such manipulations, especially in large-scale social science surveys, are usually expensive. Moreover, many well-established social science longitudinal surveys may simply not be willing to allow researchers to hold out some respondents to be interviewed only occasionally. Principal investigators may fear that such experiments interfere with their historically grown panels. In fact, we are not aware of any intentionally implemented panel conditioning experiments in large-scale longitudinal social science surveys such as the Panel Study on Income Dynamics, the German Socio-Economic Panel or the British Household Panel. Only recently, with the emergence of online panels (e.g., the Longitudinal Internet Study for the Social Sciences in the Netherlands or the German Internet Panel), have some surveys opened their panels for experimental manipulations and methodological experiments.

Variant four can overcome the difficulties discussed above because researchers rely on experimental manipulations that are already (unintentionally) implemented in surveys some other way. The drawback of this variant is that researchers can only work with what is available.

That is, while some surveys use refreshment samples or rotating designs (see above) that can be used for uncovering panel conditioning effects, others simply do not apply such designs. Thus, this variant, although very powerful in producing unbiased control group outcomes, completely depends on the availability of refreshment samples or rotation groups.

Variant two also depends on the question whether researchers can implement experimental manipulations in a panel survey. While it may be easier to implement variant two than variant one because *all* respondents are interviewed, the drawback of variant two is that researchers need to be sure that asking some other (placebo) questions does not influence the outcome in some other way.

Variant three requires somewhat different assumptions in order to produce an unbiased control group outcome. Researchers need a cross-sectional survey that asks the same question. Furthermore, this survey should be conducted in the same mode (e.g., CATI or CAPI) to avoid mixing up mode effects with panel conditioning effects (see also the discussion in Section 2.4.3). Data should be collected at the same time because attitudes and behavior may follow seasonal trends and change over time for other external reasons (see examples above). Samples of both surveys should be drawn with the same design and from the same population to avoid selection bias. Thus, in order for this variant to produce unbiased results, researchers need to find two almost identical surveys with the only difference between the two being that one is conducted as a longitudinal survey, i.e., interviews respondents more than once with the same questionnaire, and the other one as a cross-sectional survey. Given these requirements, variant three seems difficult to implement.

To sum up, the within-person approach and the four variants of the between-person approach all provide helpful designs for the analysis of panel conditioning effects. The between-person variants are more powerful in most settings as they are not restricted to certain forms of panel conditioning and may require fewer assumptions to produce unbiased control group outcomes. Both approaches (even if based on careful experimental manipulations), however, can produce biased estimates of panel conditioning if other confounding sources of error in (panel) surveys are not properly accounted for.

2.4.3 Challenge 3: Accounting for confounding sources of error

The third challenge is that researchers need to account for confounding sources of error when analyzing panel conditioning effects. The most common are longitudinal nonresponse (i.e., panel attrition), mode effects and interviewer effects.

Panel attrition

The main confounder in studies of panel conditioning is panel attrition, i.e., nonresponse in follow-up waves of a panel survey (e.g., Das, Toepoel, and van Soest 2011). Panel attrition is usually highly selective: it is related to certain observable and/or unobservable characteristics of a respondent, leading to compositional differences between all respondents of the first wave and those who also participate in subsequent waves of a panel. For example, respondents may differ from the full sample regarding socio-demographics or personality traits (e.g., Lepkowski and Couper 2002; Lugtig 2014). In these cases, attrition can easily be mistaken for panel conditioning because changes in the composition of the panel respondents over time can also affect the outcome where panel conditioning is suspected. Consider variant four of the between-person approach, for example. The distribution of socio-demographic information in the panel respondent group will be different from the distribution of these characteristics in the incoming sample due to nonrandom attrition.[7] If attrition is also correlated with the outcome, then comparisons of the mean of this variable between panel respondents and the refreshment sample could reveal (or hide) a panel conditioning effect that, in fact, is only due to nonrandom dropout over time among panel respondents. As a result, panel attrition bias would be mis-characterized

[7] Initial nonresponse is usually less of a problem because it affects both samples to the same degree. It may affect estimates of panel conditioning only if nonresponse patterns in the first interview of the treatment group are different from nonresponse patterns of the first interview of the control group. For example, nonresponse patterns in a survey of attitudes on same-sex marriage may differ between the treatment group and the control group. If same-sex marriage were to be legalized between the first wave of the treatment group and the first wave of the control group, then nonresponse patterns may differ between the two groups. Some opponents of the new bill who would have responded to the survey before the introduction of the bill may in fact not respond after legalization as they may be afraid to share their opposing opinion in an interview due to social desirability. In such cases, initial nonresponse patterns may differ between the first round of the treatment group and the first round of the control group causing an additional challenge for the analysis of panel conditioning effects. The methods that adjust for panel attrition (discussed below), however, can easily be extended to adjust for different (initial) nonresponse patterns, too.

as panel conditioning.

Studies of panel conditioning implement various approaches to tackle panel attrition. A few studies recognize attrition as a confounding source of error, but do not adjust for it in their empirical analysis (e.g., Bailar 1975) or subsume conditioning *and* attrition under 'panel bias' (e.g., Bartels 1999). Many more studies address it using one of the adjustment techniques reviewed below.

Broadly speaking, we can categorize approaches to adjust for attrition into three classes. The first class comprises approaches that simply disregard those respondents who do not participate in all waves of a panel. These approaches do not require any assumptions about the form of panel attrition and are therefore a powerful way to eliminate any bias in panel conditioning effects due to panel attrition. A drawback of these approaches, however, is that they can be implemented in combination with the within-person approach (e.g., Sturgis, Allum, and Brunton-Smith 2009, see Section 2.4.2) and variants two and four of the between-person approach only (e.g., Halpern-Manners and Warren 2012, see Section 2.4.2). Disregarding attriters cannot be used in combination with variants one and three of the between-person approach because only the treatment groups are affected by attrition. Thus, limiting the treatment groups to respondents who participate in all waves would not account for attrition at all. Instead, it would exacerbate error due to attrition because additional cases would be excluded from the group that already suffers from nonrandom dropout. Therefore, this approach is limited to certain scenarios.

The study by Halpern-Manners and Warren (2012), implementing this approach in combination with variant four of the between-person approach, deserves a closer look as the authors implement one of the most elaborate research designs to the analysis of panel conditioning effects. In the simplest case, their design consists of two rotation groups, as shown in Table 2.2. The treatment group is first interviewed in Wave 1 and interviewed for a second time in Wave 2. The control group is interviewed for the first time in Wave 2 and for the second time in Wave 3. Restricting both groups to those respondents who participate in two waves (Wave 1 and Wave 2 for the treatment group and Wave 2 and Wave 3 for the control group) and comparing the Wave 2 outcome of the treatment group (their second interview) with the Wave

Table 2.2: Example of a research design accounting for attrition with a rotating panel survey

	Wave 1	**Wave 2**	**Wave 3**
Rotation group 1	X	X	
Rotation group 2		X	X

Comparing the Wave 2 outcomes between Rotation group 1 and Rotation group 2 accounts for attrition because both rotation groups contain only respondents who participate in all waves.

2 outcome of the control group (their first interview) eliminates any confounding error due to panel attrition without having to impose any further assumptions. Thus, this design is very powerful in eliminating any confounding error due to panel attrition from panel conditioning and in providing unbiased control group data. Yet, as we have noted before, it is limited to those surveys that apply the refreshment or rotation group design described in Section 2.4.2.

The second class of approaches comprises all methods that adjust for attrition by conditioning on observable determinants of attrition (e.g., by including them as control variables in regression functions) or functions of the determinants (e.g., by using them as weights). These approaches usually assume that attrition is missing at random (MAR), i.e., attrition is random conditional on some fully observed covariates (Rubin and Little 2002). In other words, any confounding bias due to attrition in estimates of panel conditioning is adjusted for when all variables that determine whether a respondent drops out over time or not are correctly accounted for. The difficulty with this approach, however, is that the MAR assumption cannot be tested or verified in a statistical way. That is, researchers can only *assume* that attrition is MAR by, for example, conditioning on all determinants identified as predictors of attrition in previous research or all covariates thought to be related to attrition based on theory. If researchers fail to include relevant variables or if attrition is missing not at random (MNAR), then panel conditioning effects will be confounded by attrition.

Empirical implementations of this class of approaches are numerous. Kruse et al. (2009), for example, adjust for attrition by including determinants of attrition as control variables in their regression model. Other studies (e.g., Pennell and Lepkowski 1992; Dennis 2001; Nancarrow and

Cartwright 2007), by contrast, condition on determinants of attrition by including nonresponse weights supplied by the panel survey (e.g., longitudinal nonresponse weights or poststratification weights) or calculate their own weights based on propensity scores (e.g., Struminskaya 2016). Regardless, these studies cannot test whether the MAR assumption actually holds.

Approaches based on the MAR assumption are often used because they can be combined with *each* of the approaches presented in Section 2.4.2. Preference should be given to the first class of approaches (restricting the sample to respondents of all waves) however, because they do not require relying on an untestable assumption.

A third approach to account for attrition uses instrumental variables. That is, the third approach does not require researchers to assume that they observe all determinants of attrition. Instead, this method exploits random allocation of people to a treatment and a control group. We are aware of only one study that has applied this approach to account for attrition. We therefore describe the approach with the example of this study.[8] Bach and Eckman (2017) observe the behavior of two groups of people (a treatment group and a control group) in administrative records to study a changes-in-behavior panel conditioning effect. The treatment group consists of respondents and nonrespondents of a panel survey, and the control group is observed in the administrative data only. To account for attrition and initial nonresponse, the authors instrument the (endogenous) participation in the survey among members of the treatment group with the random allocation of people to the treatment or the control group. Because the allocation of people to the two groups is random and correlated with the endogenous treatment (actual participation in the survey), it is a valid instrument by definition. That is, this method exploits the random variation from the allocation of people to treatment or control to overcome the endogeneity resulting from members of the treatment group self-selecting into panel response or nonresponse. Thus, this method is a powerful tool to eliminate bias due to attrition and initial nonresponse among members of the treatment group. A major drawback of the approach, however, is that it works only when outcome information can be observed for respondents, nonrespondents, and the control group. Moreover, instrumental variables will

[8] Crossley et al. (2017) apply the same approach to account for nonresponse in their study of the effect of participating in only one wave of a panel survey on respondents' behavior.

likely underestimate the true panel conditioning effect because all respondents who participate in at least one wave of a panel have to be treated as panel respondents (see Bach and Eckman 2017, for a discussion).

All three classes of approaches to account for confounding error due to panel attrition can be used to uncover unbiased panel conditioning effects. The choice of the method, however, is often limited by the availability of data and the design applied by the panel study of interest. Even when attrition is successfully accounted for, other sources of error still need to be considered in order to estimate unbiased panel conditioning effects.

Other confounding sources of error

Mode and interviewer effects are two other sources of error in surveys that require adjustment (Halpern-Manners and Warren 2012; Chadi 2013). Both phenomena can lead to changes in measurement error in panel surveys over time and may therefore easily be mistaken for panel conditioning. Halpern-Manners and Warren (2012), for example, report that many studies of panel conditioning in the Current Population Survey (CPS) likely estimate biased conditioning effects because they confound conditioning with mode effects as the survey uses mostly in-person interviews in a respondent's first round and telephone interviews in subsequent rounds. Thus, changes in an outcome between the first round and subsequent rounds of the CPS may (partly) be due to face-to-face and telephone interviews resulting in different levels of measurement error. Halpern-Manners and Warren (2012) demonstrate, however, that the CPS is still affected by panel conditioning, even when mode effects have been properly accounted for.

Regarding interviewer effects, Chadi (2013) shows that self-reported life satisfaction decreases steadily over time due to increased trust in the interviewer (a changes-in-reporting effect). Once panel respondents are interviewed by a new interviewer, however, he finds an abrupt rise in life satisfaction (an interviewer effect). Without acknowledging changes in interviewer allocation, such abrupt rises in life satisfaction may be falsely attributed to panel conditioning. Similarly, Van der Zouwen and Van Tilburg (2001) report that changes in reported network size over time are due to interviewer behavior and not panel conditioning. Moreover, other substantial changes in panel surveys, such as a change of the data collection agency or the introduction of dependent interviewing, might lead to similar changes over time that can easily be mistaken for panel conditioning. However, we are not aware of any published research regarding these other sources of error. To sum up, researchers need to make sure that any changes found over time are only due to repeated interviewing of the same people and not due to other elements of a survey that might change between waves and cause change in respondents' behavior and/or reporting.

2.5 Discussion and a look forward

The purpose of this study was to give an overview of the current state of methodological research on panel conditioning effects in social science longitudinal surveys. We have identified three methodological challenges that have plagued research on panel conditioning for a long time. First, panel conditioning can result in changes-in-behavior and/or changes-in-reporting. To make statements about either form, it is essential to clearly disentangle the two forms (at minimum, researchers should acknowledge that panel conditioning can take various forms). Second, obtaining control group data of people who have not been interviewed or have been interviewed only once is crucial to the identification of the causal effect of panel survey participation on inter-wave changes. Third, panel attrition and other sources of error in (panel) surveys have to be accounted for to estimate unbiased panel conditioning effects. We have reviewed these challenges and the research designs developed to tackle them and discussed their strengths and weaknesses.

The discussion of these designs and their implementations in the literature have shown that more methodologically sound research is needed that carefully tackles all of the challenges identified in this study. In addition, we would like to see more studies on the kind of large-scale social science longitudinal surveys that social scientists frequently use for substantive research. For example, we are not aware of any research on panel conditioning effects in the Panel Study on Income Dynamics, the world's longest-running household panel study. However, we also acknowledge that the implementation of methodological experiments is expensive, difficult or even impossible. As a result, the choice of research designs may be limited to those ex post designs that require several (untestable) assumptions, making clear statements about the presence of panel conditioning effects difficult or even impossible.

Future research should also go beyond uncovering *average* panel conditioning effects. Little is known about treatment effect heterogeneity, i.e., how panel conditioning effects vary for different subgroups. Panel conditioning due to learning (see Section 2.3.1), for example, may vary with respondents' cognitive ability. Research should also extend the current focus on univariate panel conditioning to the multivariate context. To date, little is known about the ways panel conditioning biases estimates derived from complex multivariate statistical models. Future work should assess how panel conditioning translates into bias in, e.g., regression coefficients derived from econometric panel data models. Such projects may, for example, simulate data without panel conditioning from real data that is affected by panel conditioning. Comparing estimated panel regression coefficients between the two datasets would, for example, allow to assess bias in multivariate estimates due to panel conditioning.

Recent technological advancements in data collection techniques, such as the use of mobile (smart)phones, create even more opportunities for studies on conditioning effects. Online panel surveys, text message surveys and app-based surveys, for example, offer new means of data collection that are easy to implement, cheap and often allow to interview respondents up to several times per month (online panels) or even per week (text message and app-based surveys). At the same time, however, high-frequency surveying is also most susceptible to conditioning (see Section 2.3). Thus, when collecting high-frequency data via text message surveys, smartphones or online surveys, special attention must be paid to the possibility that the act of measuring changes what is being measured. If researchers incorporate experimental manipulations in such new data collection designs from the very beginning, obtaining sound and unbiased estimates of panel conditioning will be easy.

Any experiment on panel conditioning, however, should only be conducted if there is theoretical reason to assume a conditioning effect to be present (for example, if researchers believe that answering to a survey item encourages respondents to change their behavior). Similarly, studies should be restricted to those items routinely used in surveys because the scientific value of studying items that are hardly ever part of a survey may be questionable, a point also raised by Warren and Halpern-Manners (2012).

Furthermore, findings regarding changes-in-behavior panel conditioning may also open new ethical debates about survey research. That is, as soon as questions change behavior, researchers must carefully think about the resulting behavioral change from an ethical point of view. Veroff, Hatchett, and Douvan (1992), for example, demonstrated that asking questions about marriage may lead to decreased marital satisfaction. Similarly, asking questions about suicidal thoughts in a survey could increase the likelihood of suicide among certain respondents. Thus, some questions may have huge negative influences on respondents' behavior. Research on panel conditioning therefore may also provide guidelines for researchers collecting data regarding the ethics of asking certain questions.

Finally, no recommendations exist regarding what to do when panel data are affected by panel conditioning. Are data really "irredeemably biased" as Warren and Halpern-Manners (2012, p. 522) put it? Or is it possible to 'repair' data or account for panel conditioning by, e.g., adjusting statistical models? Can survey designers imagine ways to avoid panel conditioning in the first place? Although our review has shown that scholars have analyzed for decades how repeated interviewing of the same people can change their (reporting of) behavior and attitudes, we have also identified several challenges that many studies fail to address adequately. Similarly, our discussion shows that many important issues related to conditioning effects still need to be addressed.

References

Angel, Stefan, Richard Heuberger, and Nadja Lamei (2017). “Differences Between Household Income from Surveys and Registers and How These Affect the Poverty Headcount: Evidence from the Austrian SILC”. In: *Sociological Indicators Research* Advance online publication, DOI 10.1007/s11205-017-1672-7.

Axinn, William G., Elyse A. Jennings, and Mick P. Couper (2015). “Response of Sensitive Behaviors to Frequent Measurement”. In: *Social Science Research* 49, pp. 1–15.

Bach, Ruben L. and Stephanie Eckman (forthcoming). “Motivated Misreporting in Web Panels”. In: *Journal of Survey Statistics and Methodology.*

— (2017). “Does Participating in a Panel Survey Change Respondents’ Labor Market Behavior?” In: *IAB Discussion Paper* 15/2017. Available at `http://doku.iab.de/discussionpapers/2017/dp1517.pdf`, last accessed Jan 23, 2018.

Bailar, Barbara A. (1975). “The Effects of Rotation Group Bias on Estimates from Panel Surveys”. In: *Journal of the American Statistical Association* 70.349, pp. 23–30.

— (1989). “Information Needs, Surveys, and Measurement Errors”. In: *Panel Surveys.* Ed. by D. Kasprzyk, G. J. Duncan, G. Kalton, and M. P. Singh. New York: Wiley, pp. 1–24.

Bartels, L. (1999). “Panel Effects in the American National Election Studies”. In: *Political Analysis* 8, pp. 1–15.

Battaglia, Michael P., Elizabeth R. Zell, and Pamela L. Y. H. Ching (1996). “Can Participating in a Panel Sample Introduce Bias into Trend Estimates?” In: *Proceedings of the American Statistical Association, Survey Research Methods Section*, pp. 1010–1013.

Borle, Sharad, Uptal M. Dholakia, Siddharth S. Singh, and Robert A. Westbrook (2007). “The Impact of Survey Participation on Subsequent Customer Behavior: An Empirical Investigation”. In: *Marketing Science* 26.5, pp. 711–726.

Cantor, David (2010). “A Review and Summary of Studies on Panel Conditioning”. In: *Handbook of Longitudinal Research. Design, Measurement and Analysis.* Ed. by S. Menard. Amsterdam: Elsevier, pp. 123–138.

Chadi, Adrian (2013). “The Role of Interviewer Encounters in Panel Responses on Life Satisfaction”. In: *Economic Letters* 121, pp. 550–554.

Chandon, Pierre, Vicki G. Morwitz, and Werner J. Reinartz (2005). “Do Intentions Really Predict Behavior? Self-Generated Validity Effects in Survey Research”. In: *Journal of Marketing* 69.2, pp. 1–14.

Clausen, Aage R. (1969). “Response Validity: Vote Report”. In: *The Public Opinion Quarterly* 32, pp. 588–606.

Clinton, Joshua D. (2001). *Panel Bias from Attrition and Conditioning: A Case Study of the Knowledge Networks Panel.* Tech. rep. Stanford, CA: Department of Political Science Technical Report, Stanford University.

Cohen, Steven B. and Vicki L. Burt (1985). “Data Collection Frequency Effect in the National Medical Care Expenditure Survey”. In: *Journal of Economic & Social Measurement* 13.2, pp. 125–151.

Coombs, Lolagene C. (1973). “Problems of Contamination in Panel Surveys: A Brief Report on an Independent Sample, Taiwan, 1970”. In: *Studies in Family Planning* 4.10, pp. 257–261.

Crossley, Thomas, Jochem de Bresser, Liam Delaney, and Joachim Winter (2017). “Can Survey Participation Alter Household Saving Behavior?” In: *Economic Journal* 127, pp. 2332–2357.

Das, Marcel, Vera Toepoel, and Arthur van Soest (2007). “Can I Use a Panel? Panel Conditioning and Attrition Bias in Panel Surveys”. In: *CentER Discussion Paper Series* 2007-56.

— (2011). “Nonparametric Tests of Panel Conditioning and Attrition Bias in Panel Surveys”. In: *Sociological Methods & Research* 40.1, pp. 32–56.

Dennis, J M. (2001). “Are Internet Panels Creating Professional Respondents? A Study of Panel Effects”. In: *Marketing Research* 13.2, pp. 34–39.

Duan, Naihua, Margarita Alegria, Glorisa Canino, Thomas McGuire, and David Takeuchi (2007). “Survey Conditioning in Self-Reported Mental Health Service Use: Randomized Comparison of Alternative Instrument Formats”. In: *Health Research and Educational Trust* 42.2, pp. 890–907.

Eckman, Stephanie and Frauke Kreuter (forthcoming). “Misreporting to Looping Questions in Surveys: Recall, Motivation and Burden”. In: *Survey Research Methods.*

Eckman, Stephanie, Frauke Kreuter, Antje Kirchner, Annette Jäckle, Roger Tourangeau, and Stanley Presser (2014). “Assessing the Mechanisms of Misreporting to Filter Questions in Surveys”. In: *Public Opinion Quarterly* 78.3, pp. 721–733.

Granberg, Donald and Soren Holmberg (1992). “The Hawthorne Effect in Election Studies: The Impact of Survey Participation on Voting”. In: *British Journal of Political Science* 22.2, pp. 240–247.

Halpern-Manners, Andrew and John R. Warren (2012). “Panel Conditioning in Longitudinal Studies: Evidence From Labor Force Items in the Current Population Survey”. In: *Demography* 49.4, pp. 1499–1519.

Halpern-Manners, Andrew, John R. Warren, and Florencia Torche (2014). “Panel Conditioning in a Longitudinal Study of Illicit Behaviors”. In: *Public Opinion Quarterly* 78.3, pp. 565–590.

Hansen, Morris H., William N. Hurwitz, Harold Nisselson, and Joseph Steinberg (1955). “The Redesign of the Census Current Population Survey”. In: *Journal of the American Statistical Association* 50.271, pp. 701–719.

Holland, Paul W. (1986). “Statistics and Causal Inference”. In: *Journal of the American Statistical Association* 81.396, pp. 945–960.

Kraut, Robert E. and John B. McConahay (1973). “How Being Interviewed Affects Voting: An Experiment”. In: *The Public Opinion Quarterly* 37.3, pp. 398–406.

Kreuter, Frauke, Susan McCulloch, Stanley Presser, and Roger Tourangeau (2011). “The Effects of Asking Filter Questions in Interleafed Versus Grouped Format”. In: *Sociological Methods & Research* 40.1, pp. 88–104.

Kroh, Martin, Florin Winter, and Juergen Schupp (2016). “Using Person-Fit Measures to Assess the Impact of Panel Conditioning on Reliability”. In: *Public Opinion Quarterly* 80.4, pp. 914–942.

Kruse, Yelena, Mario Callegaro, J Michael Dennis, Charles DiSogra, Stefan Subias, Michael Lawrence, and Trevor Tompson (2009). “Panel Conditioning and Attrition in the AP-Yahoo! News Election Panel Study”. In: *Proceedings of the 64th conference of the American Association for Public Opinion Research (AAPOR)*, pp. 5742–5756.

Lazarsfeld, Paul F. (1940). "'Panel Studies'". In: *The Public Opinion Quarterly* 4, pp. 122–128.

Lepkowski, J. and M. Couper (2002). "Nonresponse in the Second Wave of Longitudinal Household Surveys". In: *Survey Nonresponse.* Ed. by R. Groves, J. Eltinge D. Dillman, and R. Little. New York: Wiley, pp. 259–271.

Lugtig, Peter (2014). "Panel Attrition Separating Stayers, Fast Attriters, Gradual Attriters, and Lurkers". In: *Sociological Methods & Research* 43.4, pp. 699–723.

Mathiowetz, Nancy A. and Tamra J. Lair (1994). "Getting Better? Change or Error in the Measurement of Functional Limitations". In: *Journal of Economic and Social Measurement* 20.3, pp. 237–262.

Morwitz, Vicki G., Eric Johnson, and David Schmittlein (1993). "Does Measuring Intent Change Behavior?" In: *The Journal of Consumer Research* 20.1, pp. 46–61.

Nancarrow, Clive and Trixie Cartwright (2007). "Online Access Panels and Tracking Research: The Conditioning Issue". In: *International Journal of Market Research* 49.5, pp. 573–594.

Neter, John and Joseph Waksberg (1964). "Conditioning Effects from Repeated Household Interviews". In: *Journal of Marketing* 28.2, pp. 51–56.

Oberski, Daniel, Antje Kirchner, Stephanie Eckman, and Frauke Kreuter (forthcoming). "Evaluating the Quality of Survey and Administrative Data through Multi-Trait Multi-Method Models". In: *Journal of the American Statistical Association.*

Pennell, Steven G. and James N. Lepkowski (1992). "Panel Conditioning Effects in the Survey of Income and Program Participation". In: *Proceedings of the American Statistical Association, Survey Research Methods Section*, pp. 566–571.

Rubin, Donald B. (1974). "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies". In: *Journal of Educational Psychology* 66.5, pp. 688–701.

— (1978). "Bayesian Inference for Causal Effects: The Role of Randomization". In: *Annals of Statistics* 6, pp. 34–58.

Rubin, Donald B. and Roderick J. A. Little (2002). *Statistical Analysis with Missing Data.* 2nd ed. New York: Wiley.

Schonlau, Matthias and Vera Toepoel (2015). "Straightlining in Web Survey Panels Over Time". In: *Survey Research Methods* 9.2, pp. 125–137.

Schündeln, Matthias (2017). "Multiple Visits and Data Quality in Household Surveys". In: *Oxford Bulletin of Economics and Statistics* Advance online publication, DOI: 10.1111/obes.12196.

Shack-Marquez, Janice (1986). "Effects of Repeated Interviewing on Estimation of Labor-Force Status". In: *Journal of Economic and Social Measurement* 14.4, pp. 379–398.

Shadish, Willian R., Thomas D. Cook, and Donald T. Campbell (2002). *Experimental and Quasi-Experimental Designs for Generalized Causal Inference.* Boston, MA: Houghton Mifflin.

Sharpe, J. Patrick and David G. Gilbert (1998). "Effects of Repeated Administration of the Beck Depression Inventory and Other Measures of Negative Mood States". In: *Personality and Individual Differences* 24.4, pp. 457–463.

Shockey, James W. (1988). "Adjusting for Response Error in Panel Surveys: A Latent Class Approach". In: *Sociological Methods & Research* 17.1, pp. 65–92.

Smith, Jennifer K., Alan S. Gerber, and Anton Orlich (2003). "Self-Prophecy Effects and Voter Turnout: An Experimental Replication". In: *Political Psychology* 24.3, pp. 593–604.

Solon, Gary (1986). "Effects of Rotation Group Bias on Estimation of Unemployment". In: *Journal of Business & Economic Statistics* 4.1, pp. 105–109.

Spangenberg, Eric and Carl Obermiller (1996). "To Cheat or Not to Cheat: Reducing Cheating by Requesting Self-Prophecy". In: *Marketing Education Review* 6.3, pp. 95–103.

Struminskaya, Bella (2016). "Respondent Conditioning in Online Panel Surveys. Results of Two Field Experiments". In: *Social Science Computer Review* 34.1, pp. 95–115.

Sturgis, Patrick, Nick Allum, and Ian Brunton-Smith (2009). "Attitudes Over Time: The Psychology of Panel Conditioning". In: *Methodology of Longitudinal Surveys.* Ed. by P. Lynn. New York: Wiley, pp. 113–126.

Toh, Rex S., Eunkyu Lee, and Michael Y. Hu (2006). "Social Desirability Bias in Diary Panels is Evident in Panelists' Behavioral Frequency". In: *Psychological Reports* 99, pp. 322–334.

Torche, Florencia, John R. Warren, and Andrew Halpern-Manners (2012). "Panel Conditioning in a Longitudinal Study of Chilean Adolescents' Substance Use: Evidence from an Experiment". In: *Social Forces* 90.3, pp. 891–918.

Traugott, Michael W. and John P. Katosh (1979). “Response Validity in Surveys of Voting Behavior”. In: *Public Opinion Quarterly* 43.3, pp. 359–377.

Van der Zouwen, Johannes and Theo Van Tilburg (2001). “Reactivity in Panel Studies and its Consequences for Testing Causal Hypotheses”. In: *Sociological Methods & Research* 30.1, pp. 35–56.

Van Landeghem, Bert (2012). “Panel Conditioning and Subjective Well-Being: Evidence from International Panel Data and Repeated Cross-Sections”. In: *SOEPPaper* No. 484.

Veroff, Joseph, Shirley Hatchett, and Elizabeth Douvan (1992). “Consequences of Participating in a Longitudinal Study of Marriage”. In: *Public Opinion Quarterly* 56.3, pp. 315–327.

Warren, John R. and Andrew Halpern-Manners (2012). “Panel Conditioning in Longitudinal Social Science Surveys”. In: *Sociological Methods & Research* 41.4, pp. 491–534.

Waterton, Jennifer and Denise Lievesley (1989). “Evidence of Conditioning Effects in the British Social Attitudes Panel”. In: *Panel Surveys*. Ed. by D. Kasprzyk, G. J. Duncan, G. Kalton, and M. P. Singh. New York: Wiley, pp. 319–339.

Williams, Patti, Lauren G. Block, and Gavan J. Fitzsimons (2006). “Simply Asking Questions about Health Behaviors Increases both Healthy and Unhealthy Behaviors”. In: *Social Influence* 1.2, pp. 117–127.

Wilson, Sven E. and Benjamin L. Howell (2005). “Do Panel Surveys Make People Sick? US Arthritis Trends in the Health and Retirement Study”. In: *Social Science & Medicine* 60, pp. 2623–2627.

Yalch, Richard F. (1976). “Pre-Election Interview Effects on Voter Turnout”. In: *Public Opinion Quarterly* 40.3, pp. 331–336.

Yan, Ting and Stephanie Eckman (2012). “Panel Conditioning: Change in True Value versus Change in Self-Report”. In: *Proceedings of the American Statistical Association, Survey Research Methods Section*, pp. 4726–4736.

Zwane, Alix Peterson, Jonathan Zinman, Eric van Dusen, William Pariente, Clair Null, Edward Miguel, Michael Kremer, Dean S. Karlan, Richard Hornbeck, Xavier Giné, Esther Duflo, Florencia Devoto, Bruno Crepon, and Abhijit Banerjee (2011). “Being Surveyed Can Change

Later Behavior and Related Parameter Estimates". In: *Proceedings of the National Academy of Sciences* 108.5, pp. 1821–1826.

Chapter 3

Does the Participation in a Panel Survey Change Respondents' Behavior?

3.1 Abstract

Panel survey participation can bring about unintended changes in respondents' behavior and/or their reporting of behavior. Using administrative data linked to a large panel survey, we analyze whether the survey brings about changes in respondents' labor market behavior. We estimate the causal effect of panel participation on the take-up of federal labor market programs using instrumental variables. Results show that panel survey participation leads to an increase in respondents' take-up of these measures. These results suggest that panel survey participation not only affects the reporting of behavior, as previous studies have demonstrated, but can also alter respondents' actual behavior.

3.2 Background

Panel surveys are a key resource for researchers and policy makers who seek to understand dynamic processes, such as movements in and out of the labor force. Yet such surveys are also vulnerable to the critique that participation can distort respondents' behavior and/or responses, making the collected data unrepresentative of the larger population. This phenomenon is referred to as *panel conditioning* (Halpern-Manners, Warren, and Torche 2017). Although concerns about panel conditioning first arose in the 1940s (Lazarsfeld 1940), researchers from many disciplines still rely on panel data for causal analysis. In this study, we test whether repeated participation in the large-scale German panel study "Labor Market and Social Security" alters respondents' labor market behavior. We think of participation in the first three waves of the panel survey as a treatment and the panel conditioning effect as a treatment effect. The outcome variables of interest are take-up of federal labor market programs and job search behavior. That is, we use techniques of causal analysis to study whether panel survey participation makes respondents more or less likely to take part in the labor market programs and whether it helps respondents find a job faster.

Our study faces two major methodological challenges. First, we need to disentangle the effect of survey participation on changes-in-behavior from changes-in-reporting. Administrative labor market data, which are independent of respondents' reporting, make this disentanglement possible. Second, we need to control for confounding effects of survey nonresponse and attrition from the estimates of panel conditioning. We use an instrumental variable approach and select a second sample of persons who were eligible for selection into the panel survey but were not selected. Thus, our data consist of two random subsamples – one selected for the survey, the other not. Instrumenting actual participation in several waves of the survey, i.e., the treatment, with the (random) invitation to participate in the survey, we adjust for bias due to nonresponse and attrition and estimate the causal effect of panel survey participation on respondents' labor market behavior.

Previous studies of behavioral changes due to panel survey participation are rare and most suffer from design flaws. Many do not disentangle changes-in-reporting from changes-in-behavior due to data limitations; others do not untangle panel conditioning effects from other effects such as attrition. We contribute to this sparse literature and tackle the methodological challenges that previous studies have encountered.

Before we turn to the data, methods and results of our study, we discuss panel conditioning and its two forms in more detail. We also explain why we expect panel participation to alter survey respondents' labor market behavior.

3.2.1 Panel conditioning

If participation in a panel survey induces changes-in-behavior, then the survey's sample becomes less representative of the population over time, and estimates based on the data will be biased (Yan and Eckman 2012). The classic finding of this type of panel conditioning, from the field of political science, is that participation in a pre-election survey increases voter turnout in upcoming elections (Clausen 1969; Kraut and McConahay 1973; Yalch 1976; Traugott and Katosh 1979; Granberg and Holmberg 1992). However, not all studies have detected this effect (Smith, Gerber, and Orlich 2003). Panel participation can also affect other types of behavior: water treatment product use and purchases of health insurance (Zwane et al. 2011); purchases of automotive services (Borle et al. 2007), automobiles (Morwitz, Johnson, and Schmittlein 1993; Chandon, Morwitz, and Reinartz 2005) and computers (Morwitz, Johnson, and Schmittlein 1993); saving for retirement (Crossley et al. 2017)); and cheating in exams (Spangenberg and Obermiller 1996). For an extensive review of relevant studies in consumer behavior and marketing research, see Dholakia (2010).

The literature proposes several theoretical explanations for changes-in-behavior panel conditioning (Warren and Halpern-Manners 2012). Two mechanisms best explain why we expect changes in behavior due to panel conditioning to arise in our labor market survey data. Cognitive stimulus theory holds that repeatedly being asked the same questions makes respondents more aware of the topic of the survey, raises their consciousness of the issues, and motivates them to engage in the behavior under study (Sturgis, Allum, and Brunton-Smith 2009; Zwane et al. 2011; Warren and Halpern-Manners 2012). This approach likely explains panel conditioning in studies of voting behavior: a pre-election interview may stimulate interest and participation in the election and thereby increase voter turnout among respondents (Clausen 1969). The second explanation is concerned with stigmatized or socially non-normative behavior. When survey questions force respondents to confront a conflict between their behavior and society's norms, they might bring their future behavior in line with social norms to avoid dissonance (Williams, Block, and Fitzsimons 2006).

Based on these two theoretical explanations, we hypothesize that repeatedly answering questions in a panel survey about whether or not one has participated in active labor market policies (ALMP) increases the likelihood that respondents will participate in those measures. ALMP are programs administered by the German government that aim to reduce unemployment and increase participation in the labor market (Crépon and Van den Berg 2016): they may include measures such as job application trainings or continuing education courses. Participation in such measures is often mandatory for recipients of unemployment benefit recipients, and failure to participate may result in sanctions, i.e., temporary cuts in benefit receipt (Jacobi and Kluve 2007). Taking part in a survey that includes several questions about such programs may stimulate respondents to think more about the ALMPs, enhance their awareness of them and thereby increase the likelihood that respondents will participate in these ALMPs. Additionally, respondents may feel embarrassed reporting that they have not participated in such measures and change their future behavior accordingly. In this way, panel participation could change some respondents' behavior. Furthermore, we expect that the size of the effect of panel survey participation on respondents' labor market behavior will increase with each wave.

Participating in the panel may also increase respondents' likelihood to find employment. That is, because ALMP are designed to help unemployment benefit recipients find a new job, we expect that participating in the panel also reduces the search time that respondents need to find a job. It is also possible that survey participation has an independent effect on job search success, perhaps due to the stigma explanation discussed above.

Panel conditioning can also induce changes in how respondents report their behavior, which we refer to as changes-in-reporting (Waterton and Lievesley 1989; Sturgis, Allum, and Brunton-

Smith 2009; Cantor 2010; Warren and Halpern-Manners 2012). For example, respondents may report more accurately in later waves of a panel survey, due to increased trust in the interviewing process. On the contrary, respondents may also learn from prior participation how the questionnaire is structured and then falsify their answers to reduce the length of the interview. For a detailed review of theoretical explanations for changes-in-reporting panel conditioning and an extensive literature review see, e.g., Warren and Halpern-Manners (2012).

Both forms of panel conditioning can occur at the same time (Halpern-Manners and Warren 2012; Yan and Eckman 2012). For example, unemployed respondents of a panel survey on labor market behavior may under-report unemployment due to social desirability bias in early waves of the survey but report more truthfully in later waves as they become more trustful of the interviewer. Yet, at the same time, repeatedly asking respondents about labor market topics could stimulate respondents' job search activities and thereby lead to changes in their actual behavior (the cognitive stimulation hypothesis). Researchers need to be aware of the various ways that panel conditioning can occur and how it can bias inference: working with data affected by panel conditioning, researchers risk mischaracterizing the existence, magnitude and correlates of changes across survey waves in respondents' attitudes and behaviors, which are the main estimates made from panel data (Clinton 2001; Halpern-Manners, Warren, and Torche 2017). In addition, as mentioned above, the conclusions drawn from the panel participants may not generalize to the larger population.

3.2.2 Methodological challenges

Studies of panel conditioning are confronted with two major methodological challenges. The first is that disentangling changes-in-behavior from changes-in-reporting is difficult or impossible to do without validation records for the answers given in the survey (Waterton and Lievesley 1989; Van der Zouwen and Van Tilburg 2001). Indeed, most studies of changes-in-reporting do not distinguish between the two types of panel conditioning (for exceptions see Pennell and Lepkowski (1992), Van der Zouwen and Van Tilburg (2001), Duan et al. (2007), and Halpern-Manners and Warren (2012)). To isolate changes-in-behavior panel conditioning, we need data that are unaffected by respondents' reporting, such as administrative process data. Using such data, researchers can study changes-in-behavior by, for example, comparing respondents' behavior in the administrative data with the behavior of those who were not interviewed. This is the approach we follow in our study. If one was interested in changes-in-reporting, comparisons of respondents' survey answers with administrative records may be used (e.g., Yan and Eckman 2012). If only survey data are available, one may compare experienced respondents with novel respondents in panel studies with rotating designs to study changes-in-reporting, but this approach does not fully separate between the two forms of panel conditioning (Halpern-Manners and Warren 2012).

The second major challenge to estimating panel conditioning error is adjustment for confounding sources of error in panel studies (Williams and Mallows 1970; Van der Zouwen and Van Tilburg 2001; Sturgis, Allum, and Brunton-Smith 2009; Das, Toepoel, and van Soest 2011; Halpern-Manners and Warren 2012). The most common confounding source of error is nonresponse and panel attrition. Initial nonresponse, for example, will result in systematic differences between respondents of a survey and the originally selected sample. If these differences are not properly adjusted for, analysis of panel conditioning may be confounded by nonresponse bias. Similarly, compositional changes in the survey sample due to nonresponse in later waves of a panel survey, i.e., panel attrition, may be mistaken for panel conditioning. For example, almost all studies of panel conditioning in the Current Population Survey (CPS) do not distinguish between attrition and conditioning (Halpern-Manners and Warren 2012). Some researchers, using other data sets, do attempt to control for the effects of nonresponse and attrition by conditioning on covariates related to nonresponse and attrition; others attempt to exclude attrition by comparing only those who did not attrit (Shack-Marquez 1986; Pennell and Lepkowski 1992; Dennis 2001; Nancarrow and Cartwright 2007; Toepoel, Das, and van Soest 2009; Das, Toepoel, and van Soest 2011; Halpern-Manners and Warren 2012). Two other potential sources of error that can easily be mistaken for panel conditioning are interviewer effects and mode effects. Van der Zouwen and Van Tilburg (2001), for example, show that changes in reported network size over time are due to interviewer changes between two waves of a panel survey and not panel conditioning. Similarly, Halpern-Manners and Warren (2012) warn that changes over time may also result from a change of the data collection mode: the CPS, for example, uses in person interviews in the first and telephone interviews in subsequent rounds.

We note that, among large socio-economic panel surveys, only one study of changes-in-behavior does not suffer from the problems discussed above. Crossley et al. (2017) employ a quasi-experimental design to identify a survey participation effect in a large-scale panel survey in the social sciences. Analyzing administrative wealth data from respondents of a Dutch online panel, they find that participating in an interview about saving for retirement has, on average, a negative effect on respondents' future saving behavior. Although the authors estimate the effect of participation in a survey module fielded only once, their results lend support to the hypothesis that (repeated) survey participation can alter respondents' behavior. Moreover, their identification strategy, which uses an instrumental variable approach, can be applied to the estimation of panel conditioning effects, and that is the approach we take in this study.

3.3 Data

Data for our analysis come from the German panel study "Labor Market and Social Security" (PASS), which is a large-scale yearly panel survey conducted by the Institute for Employment Research on labor market topics. It uses a mixed-mode design of computer-aided telephone interviews (CATI) and computer-aided personal interviews (CAPI). The survey consists of a household interview and additional individual interviews with all household members who are at least 15 years old. The main topics of the interviews are pathways into and out of unemployment benefits receipt type II (long-term benefits due to unemployment, disability or employment that does not reach a minimum standard of living); dynamics of the material and social situation of benefit recipients; changes in recipients' behavior and attitudes over time; and interactions between recipients and the benefit providing agencies (Trappmann et al. 2013).

The PASS sample consists of two subsamples. The *recipient subsample* ($n = 29,309$) is a representative sample of all unemployment benefit units in Germany drawn from a register maintained by the Federal Employment Agency. The sample was drawn from the most recent available administrative records in July 2006. Unemployment benefit units are in most cases identical to households (see Trappmann et al. 2013). Variables on the register, however, concern the individuals residing in these households. The second subsample is a *general population sample* of households, selected from a commercial data set of addresses in Germany. We do not use this subsample in our analysis and do not describe it further (see Trappmann et al. 2013, for details).

Both samples were drawn using a multi-stage sampling design (Schnell 2007; Rudolph and Trappmann 2007). In the first stage, 300 postcode areas were selected as primary sampling units. In the second stage, unemployment benefit units in the selected postcode areas were drawn from the administrative register of unemployment benefit recipients. Since the recipient sample was drawn from the administrative data, respondents and nonrespondents can easily be identified in the administrative records. We discuss linkage between the survey and administrative data in more detail below.

We consider data from the first three waves of PASS, excluding refreshment samples that were introduced in several waves (Trappmann et al. 2013). Data were collected on an annual basis between winter 2006 and spring 2009. In each of these waves, respondents were asked the same set of questions about their participation in ALMPs (see Appendix A for the text of the questions). These questions may provide a cognitive stimulus that will increase respondents' awareness and take-up of the programs, and for this reason we hypothesize changes-in-behavior panel conditioning.

Household response rates are 35 percent in Wave 1, 51 percent in Wave 2, and 65 percent in Wave 3 (Wave 2 and Wave 3 response rates conditional on participation in previous waves, Christoph et al. 2008; Gebhardt et al. 2009; Berg et al. 2010). However, in our dataset we cannot distinguish between households selected and fielded ($n = 23,736$) and households selected and not fielded ($n = 5,573$). Therefore, the household response rate calculated with our larger dataset is 22.8 percent. Since the subset of households used during fieldwork is a random subsample of the original sample, we do not expect the additional households to bias our estimates.

3.3.1 Administrative data and linkage

The administrative data we use to investigate changes-in-behavior conditioning are the 'Integrated Employment Biographies' (IEB), which include records of all employment spells subject to social security, unemployment benefit receipt, participation in ALMP, or spells of job search (Jacobebbinghaus and Seth 2007; Institute for Employment Research 2013). These records can be aggregated to the person level and contain histories of employment, unemployment, job search, and benefit receipt, as well as records of ALMP participation. Although the administrative data are not free of error, their overall quality is very high, and they are used by the German government to calculate pension claims, administer benefit claims and make payments (Jacobebbinghaus and Seth 2007; Köhler and Thomsen 2009; Kreuter, Müller, and Trappmann 2010). In general, data on benefits, ALMP and job search are of the highest quality, because they are generated by activities of the Federal Employment Agency itself (Jacobebbinghaus and Seth 2007). A variety of socio-demographic variables such as gender, date of birth, citizenship, education and place of residence are also included in the data set.

Because the PASS recipient sample is selected from the same data source, we can easily identify members of the households sampled for PASS in the IEB. Overall, we are able to identify and obtain full information on 28,377 selected households, i.e., 96.8% of the original household sample. Cases that do not have socio-demographic variables needed for our analysis (see Section 3.4.5) are not included (3.2%). We do not expect this restriction of the sample to bias our estimates of panel conditioning because households with similar scarce administrative records will also not be included in our control group, for the same reason (see Section 3.4). For the 28,377 households with full information, we obtain individual records on 38,350 household members.

Table 3.1 shows the number of households with at least one realized interview per wave and the number of households successfully identified in the administrative records. Overall, we are able to identify and obtain full information on 6,415 households that responded in at least one of the first three waves of PASS (95.3%, last column of Table 3.1). Likewise, we obtain full information on 21,962 households that did not respond to any of the first three waves of PASS

Table 3.1: Number of households with at least one realized interview and number of successfully linked households per wave

	Wave 1	**Wave 2**	**Wave 3**	**Response in any wave**
Households with at least one realized interview	6,691	3,391	3,615	6,734
Responding households successfully linked to administrative records	6,371 (95.2%)	3,236 (95.4%)	3,445 (95.3%)	6,415 (95.3%)
Individuals within linked responding households	8,647	4,337	4,626	8,728

(97.3%). Thus, the linkage rate among all households is very high.

Due to German privacy regulations, we cannot link individual response indicators from the survey to the administrative data without respondents' consent. For this reason, we treat all members of a responding household as respondents. Restricting our analysis sample to consenting respondents may introduce bias if consenters are different from non-consenters. To avoid such bias, we do not identify individuals within households and treat *all* household members as respondents if at least one person in the household responded to the survey. This issue might lead us to underestimate the true panel conditioning effect because some household members have not received the stimulus of responding to the survey (within-household response rates at the individual level are 85.6% in Wave 1, 85.5% in Wave 2, conditional on response in Wave 1, and 83.5% in Wave 3, conditional on response in Wave 2, Christoph et al. 2008; Gebhardt et al. 2009; Berg et al. 2010). However, it avoids any bias due to non-consent in our estimates. Based on these definitions, about 22.8 percent of all individuals identified in the administrative data responded to at least one wave of PASS.

Identifying the administrative records for respondents and nonrespondents is only the first step, however: to get an unbiased estimate of the effect of repeated survey participation on ALMP take-up, we need to control for the confounding effects of nonresponse and panel attrition.

3.4 Methods

We regard participation in the first three waves of PASS as a treatment, and the panel conditioning effect as a treatment effect: we are interested in estimating how receiving the treatment (participating in three waves of the PASS survey) changes respondents' behavior. The three-time PASS respondents form the treatment group. To estimate the treatment effect, we selected a second random sample from the IEB administrative data set ($n = 38,350$) to form the control group. This sample consists of unemployment benefit recipients who were eligible for the first wave of the PASS survey but were not selected in the first wave nor in any later waves. Thus, our data consist of two random samples, one selected for the survey, i.e., the respondents and nonrespondents of PASS, the other one, the control group, not.

In formal terms, $Z \in [0, 1]$ defines an indicator of whether one was *assigned* to treatment ($Z = 1$) or control ($Z = 0$), i.e., selected for the survey or not. Moreover, $D \in [0, 1]$ denotes the treatment indicator, i.e., whether one actually *participated* in the survey ($D = 1$) or not ($D = 0$). If everyone complied with the treatment status assigned, then $D = Z$. Assignment to treatment (Z) is fixed over time. Actual treatment (D), however, may change over time, e.g., a household that did not participate in the first wave ($D = 0$) may participate in the second wave, thus changing to $D = 1$. Let Y be the outcome of interest (take-up of ALMP and job search behavior). If everyone complied with the treatment, the treatment effect would simply be the difference between the average outcomes of the treated persons $E(Y|D = 1)$ and the control group $E(Y|D = 0)$.

However, as shown in Section 3.3, many persons selected for the survey did not participate: $Z = 1$, but $D = 0$. These persons are *non-compliers*, that is, they did not respond to the survey, their assigned treatment status. Nonresponse and attrition in surveys have many causes and can lead to bias or endogenous selection into treatment (Groves, Cialdini, and Couper 1992; Groves, Singer, and Corning 2000; Abadie 2003; Kreuter, Müller, and Trappmann 2010). For example, people who agree to participate in three waves of a survey may be more compliant and thus more likely to participate in ALMP, even without the treatment of the survey (e.g., Zabel 1998; Rizzo, Kalton, and Brick 1996; Lepkowski and Couper 2002). Thus, comparing the outcomes of the treated cases with the outcomes of the control group may bias our analysis of panel conditioning, because survey respondents are different from nonrespondents in many ways.

One method to address the problem of noncompliance with treatment assignment is to estimate an intention-to-treat (ITT) effect. In this approach, rather than comparing individuals with different treatment statuses, we compare those assigned to different treatments, conditional on a vector of predetermined covariates X:

$$ITT = E[Y|Z = 1, X] - E[Y|Z = 0, X] \tag{3.1}$$

Since Z was randomly assigned, the ITT estimates the causal effect of the offer of treatment. However, due to noncompliance with the treatment status assigned (the take-up rate of the treatment, $E(D|Z = 1)$, across all waves is about 22.8 percent), the effect estimated via Equation 3.1 will be too small relative to the average causal effect on the treated (Angrist and Pischke 2009, ch. 4). A more powerful class of methods uses the randomization of Z in an indirect way to adjust for the bias due to noncompliance and to estimate the treatment effect.

3.4.1 Instrumental variables

The instrumental variable (IV) approach is based on the following idea: if an instrument Z is available that induces exogenous variation in the treatment variable D, then instrumenting D with Z allows us to estimate the treatment effect of D (Imbens and Angrist 1994; Angrist, Imbens, and Rubin 1996; Abadie 2003).

Following the potential outcomes framework (Rubin 1974; Rubin 1977) , we define two potential outcomes: Y_1 is the outcome that occurs when a case receives treatment (participates in three waves) and Y_0 is the outcome without treatment. Obviously, we observe only $Y = DY_1 + (1 - D)Y_0$ for a given individual, i.e., either Y_1 or Y_0. Furthermore, let D_z represent the potential treatment status given $Z = z$. If, for a given case $D_1 = 0$, then that case would not participate if selected; $D_1 = 1$ means that a case would participate if selected. Analogous to the potential outcomes setup, we observe only $D = ZD_1 + (1 - Z)D_0$, but never both potential treatments for any individual.

Following Angrist, Imbens, and Rubin (1996), we divide the population into four groups:

- *Compliers*: $D_1 > D_0$ or, equivalently, $D_0 = 0$ and $D_1 = 1$
- *Always-takers*: $D_1 = D_0 = 1$
- *Never-takers*: $D_1 = D_0 = 0$
- *Defiers*: $D_1 < D_0$ or equivalently $D_0 = 1$ and $D_1 = 0$

In our framework, the group of survey respondents are the compliers: They were assigned to take the treatment, i.e., selected for the survey, and complied with the treatment assignment, i.e., responded to the survey. There are no always-takers, i.e., people who take the treatment irrespective of their treatment assignment status, since participation in the survey is only possible for cases selected for participation. However, we do have never-takers, people who do not participate in the survey when they are selected (and also when they are not selected). Non-compliance in our setup is one-sided because people not assigned to the survey cannot decide between response and nonresponse. In other words, the probability that a case assigned to control does not take the treatment equals one ($Pr(D_0 = 0) = 1$). There are no defiers for the same reason.

Angrist, Imbens, and Rubin (1996) show that instrumenting D with Z estimates the local average treatment effect (LATE) for compliers under certain assumptions that we discuss in the next section. Moreover, since there are no always-takers and no defiers, the group of compliers and the group of treated are identical, and the LATE for compliers equals the average treatment effect on the treated (ATT).

3.4.2 Identification assumptions

To state the assumptions needed for the instrumental variable approach, we need to include Z in the definition of potential outcomes. Let Y_{zd} represent the potential outcome if $Z = z$ and $D = d$, and let X be a vector of known characteristics. Then, with the following nonparametric assumptions, we can use IV techniques to estimate the LATE for compliers.

(i) Independence of the instrument: conditional on X, the random vector $(Y_{00}, Y_{01}, Y_{10}, Y_{11}, D_0, D_1)$ is independent of Z.

(ii) Exclusion of the instrument: $Pr(Y_{1d} = Y_{0d}|X) = 1$ for $d \in$\{0,1\}

(iii) First stage: $0 < Pr(Z = 1|X) < 1$ and $Pr(D_1 = 1|X) > Pr(D_0 = 1|X)$

(iv) Monotonicity: $Pr(D_1 \geq D_0|X) = 1$

Assumption (i) means that treatment assignment Z is ignorable or as good as randomly assigned, conditional on X. This assumption is justified in our study. The cases selected for the PASS survey and the control data set are equal size random samples of people registered as unemployment benefit recipients in the IEB at the date of sample selection of PASS. As we would expect from random assignment of Z, there are only few significant correlations between Z and a set of covariates derived from the administrative data (see Section 3.4.5 for details on these covariates and the Appendix for supporting analysis).

Assumption (ii) states that variation in Z, i.e., assignment to treatment or control, does not influence potential outcomes except through D, i.e., through response or nonresponse. Moreover, the assumption allows us to define potential outcomes in terms of D alone, i.e., $Y_0 = Y_{00} = Y_{10}$ and $Y_1 = Y_{01} = Y_{11}$.Taken together, (i) and (ii) guarantee that the only effect Z has on Y is through D, that is, that being selected for the survey does not affect participation in ALMP except through taking part in the PASS survey.

However, we need to discuss these two assumptions in more detail. We would like to estimate the effect of repeated participation in PASS on the take-up of ALMPs. Thus, as defined above,

D refers to *repeated* participation in the survey. However, some persons participate in one or two waves and then drop out of the survey. These households are selected ($Z = 1$) but not treated ($D = 0$), because we have defined D as participation in all three waves. If participation in just one or two waves is enough to change behavior (as suggested by Crossley et al. 2017), then we have a violation of the assumptions: Z can affect Y, even when $D = 0$. Thus, we need to redefine our treatment to satisfy the assumptions. Instead of participation in the first three waves of PASS, we define the treatment as participation in any of the first three waves of PASS.

This modification to the definition of treatment strengthens our argument that, once we control for the small differences between the group selected for survey participation and the control group of unselected recipients, assignment to survey participation, Z, has no effect other than through actual participation in any wave of the survey, D, and thus that Assumptions (i) and (ii) hold. It may still be possible that being selected for the survey, receiving the advance letter announcing the survey, and perhaps receiving a few contact attempts could influence ALMP behavior without survey participation. However, we believe the chances of that happening are very small, and we are not aware of any literature showing that survey contact attempts and advance materials change behavior.

Assumption (iii) states that D and Z must be correlated, conditional on X. In addition, the support of X conditional on $Z = 1$ must coincide with the support of X conditional on $Z = 0$. Since 8,728 people (about 22.8 percent) responded in at least one of the first three waves of PASS, and no non-selected cases participated, this assumption appears to be satisfied. In addition, tests indicate that we do not need to worry about weak identification (F-statistic from an IV regression with a linear first stage: $F_{1,76678} = 11{,}187.81$; partial R-squared of the treatment indicator 0.13; Stock and Yogo (2005), Angrist and Pischke (2009, ch. 4)).

Assumption (iv), i.e., monotonicity, holds trivially because only people selected for the survey could participate in the survey.

In addition to the four assumptions discussed above, we need to assume the stable unit treatment value assumption (SUTVA) (Rubin 1978; Angrist, Imbens, and Rubin 1996). It holds that potential outcomes for each treated case are unrelated to the treatment status of others. This assumption would be violated if, for example, a responding household would interact with a non-responding household. Given the number of households with benefit receipt in Germany in July 2006 (about 4.01 million Bundesagentur für Arbeit 2017) and the sample size of PASS (29,309), the chances that households would interact are very small. Thus, we do not see any reason why this assumption would not hold.

In general, these assumptions are similar to the assumptions needed for the popular Wald estimator. In our case, however, the assumptions are more flexible as they allow conditioning on X. Since there are minor differences between the group offered to participate in the survey (see above and the Appendix for supporting analysis) and the control group, we choose the estimator presented below and control for X.

Having discussed the assumptions in detail, we next turn to the identification of the LATE or, in the terminology of Abadie (2003), the local average response function (LARF).

3.4.3 Identification

Define the object of interest, the LARF for compliers, as $E[Y|D, X, D_1 > D_0]$. The challenge for the estimation of this LARF is the identification of the compliers (those who respond when offered to participate and do not respond when not offered). Among the cases offered to participate in the survey, the compliers are easy to identify, of course: these are the respondents. In the control group, however, we cannot identify compliers individually because we cannot distinguish whether a case complies with being assigned to the control group condition or is a never-taker. Using Abadie's κ and Theorem 3.1 of Abadie (236-237 2003), however, we can express expectations for compliers with the LATE assumptions in the following way (see Section 3.4.2). Let $g(\cdot)$ be any measurable real function of (Y, D, X) with finite expectation. Furthermore, define

$$\kappa = 1 - \frac{D(1-Z)}{Pr(Z=0|X)} - \frac{(1-D)Z}{Pr(Z=1|X)} \tag{3.2}$$

The second term of κ is > 0 only for the unselected always-takers (in our cases, the second term equals 0, because people who were not offered to participate in the survey could not participate, i.e., there are no unselected always-takers). The third term is > 0 only for the selected never-takers. Thus, the only cases with $\kappa > 0$ are the compliers and the selected always-takers because there are no defiers by definition (see Section 3.4.1). Therefore, we can think of kappa as a case-level weight that identifies the complier population, allowing us to estimate the LARF for compliers. Abadie shows that $E(\kappa) = Pr(D_1 > D_0)$. Interestingly, this procedure simplifies to the the popular Wald estimator when X contains a constant only (Angrist and Pischke 2009, ch. 4). Since $Pr(Z = 1|X)$ in Equation 3.2 is unknown due to X, we estimate it, e.g., with a probit model, i.e., $Pr(Z = 1|X) = \Phi(x_i'\gamma)$, where $\Phi(\cdot)$ is the cumulative distribution function of a standard normal distribution.

3.4.4 Estimation

When Y is continuous, estimation of the LARF, $E[Y|D, X, D_1 > D_0]$, is straightforward with linear regression. Suppose $E[Y|D, X, D_1 > D_0] = h(D, X, \theta)$, with $h(D, X, \theta) = \alpha d + x'\beta$ and parameter vector $(\theta = \alpha; \beta)$. Using the results from above, Abadie (2003) proposes the following least squares estimator for continuous outcomes:

$$(\hat{\alpha}, \hat{\beta}) = arg \min_{(\alpha,\beta)\in\theta} \frac{1}{n} \sum_{i=1}^{n} \kappa_i (y_i - \alpha d_i - x_i'\beta)^2 \tag{3.3}$$

When $Pr(Z = 1|X)$ is estimated with least squares, Equation 3.3 produces the traditional two-stage least squares estimator (Angrist and Pischke 2009, ch. 4).

When the outcome, Y, is binary, a probit transformation of the linear part of (2) can be used. The estimator is then:

$$(\hat{\alpha}, \hat{\beta}) = arg \min_{(\alpha,\beta)\in\theta} \frac{1}{n} \sum_{i=1}^{n} \kappa_i (y_i - \Phi(\alpha d_i - x_i'\beta))^2 \tag{3.4}$$

In both cases, variances are estimated according to Theorem 4.2 of Abadie (2003).

Both estimators are implemented in the LARF package (version 1.4) by An and Wang (2016) in R (R Core Team 2017). We next define the outcome variables, participation in ALMP and job search after participation in the survey, and the set of covariates X in more detail.

3.4.5 Outcomes & Control variables

We consider two variables related to ALMP participation as outcomes. The first is the number of ALMP taken. The second is a simple indicator of whether a person participated in any ALMP (1) or did not (0). If panel conditioning has led to changes in behavior, we should see that respondents participate in more programs and are more likely to participate in ALMP. In building these outcome variables, we consider only those spells that started after the first day of fieldwork (because the survey cannot affect ALMP spells before it started) and those that occurred before January 31st, 2010, the day before fieldwork of wave four started (because later spells may have been influenced by later waves of the survey and in this study we consider only the first three waves). Furthermore, we consider both outcomes at three different periods in time. The first period begins after the first day of field work of Wave 1 and ends just before the beginning of Wave 2. The second period begins after Wave 2 and ends before Wave 3 and the third, finally, after Wave 3 and before Wave 4. However, we must then also define the treatment accordingly (see Section 3.4.2). Table 3.2, showing possible response patterns, defines the treatment indicator for each outcome period.

Estimating the treatment effect for each of these treatments separately allows us to study whether panel conditioning effects get stronger over time, i.e., whether the effect size increases with each additional wave. However, our approach may underestimate the true panel conditioning effect because we treat all members of a responding household as respondents (Section 3.3.1) and because the definitions of the treatments in waves two and three also include people who responded in only one wave (Section 3.4.2).

Regarding the question of whether panel survey participation increases the probability of em-

Table 3.2: Treatment variable definitions and response patterns

		Response to survey				
Outcome period	**Treatment**	**Wave 1**	**Wave 2**	**Wave 3**	**N**	**Total**
After wave one & before wave two	D=1	X			8,647	8,647
After wave two & before wave three	D=1	X	X		4,310	8,674
		X			4,337	
			X		27	
After wave three & before wave four	D=1	X	X	X	3,295	8,728
		X	X		1,015	
		X		X	1,259	
			X	X	18	
		X			1,683	
			X		9	
				X	54	

ployment, we again consider two outcome measures. First, we simply calculate the duration from the beginning of the fieldwork of the survey until the occurrence of the first job spell in the records (in days). Second, we create an indicator of whether a person found a job after the beginning of the fieldwork of the survey or not. Both variables consider only data until the beginning of wave four. For brevity, we do not consider these two outcomes at different points in time.

For each outcome, we first estimate the ITT described in (3.1). Second, we estimate the LARF with the estimator defined in (3.3) for the number of ALMP participations and the time to find a job and the estimator defined in (3.4) for the two binary participation indicators.

As a falsification or "placebo" test, we create the same two ALMP outcome variables for earlier spells of program participation, those that ended before the first day of fieldwork, as suggested by Athey and Imbens (2017). If the assumptions discussed above are met and the model is correctly specified, we should not find a significant treatment effect on pre-treatment outcomes, because the survey cannot affect ALMP spells that occurred before the survey started. We do not perform this test for the job search outcomes because it is not clear how we should define the search time for the last hob held before the survey started. In addition, most of our cases were unemployed at the time of selection and had been unemployed for a while.

Furthermore, we derive a set of covariates, X, from the administrative data. These covariates cover socio-demographics (age, gender, education, nationality, place of residence) as well as labor market histories (past employment, unemployment and unemployment benefit receipt). We include these covariates in all of our analyses, for two reasons. First, to control for any differences that might arise by chance between the group assigned to treatment and the group assigned to control. Second, because including covariates, even if they do not differ between $Z = 1$ and $Z = 0$, can reduce some of the variability in the outcome variable and thereby increase the precision of the estimates (Angrist and Pischke 2009, ch.4). For a complete list of the covariates we use, see the Appendix.

In sum, with the methods discussed here, we can estimate the causal effects of repeated participation in PASS on the take-up of ALMP and employment. The results allow us to answer our research question: whether panel participation leads to changes in respondents' labor market behavior.

3.5 Results

In a first step, we present descriptive statistics of the outcome variables. Then, we present the results of the ITT analysis and our main results, the estimated treatment effects of participation in several waves of the survey. At each step, we also check the results from the falsification test: if our models are working well, they should not detect any treatment effect in the ALMP variables before Wave 1, that is, before treatment began.

3.5.1 Descriptive statistics

Table 3.3 shows descriptive statistics of the outcome variables at four different periods in time. The first set of rows show that the two ALMP outcomes do not differ much between the group selected for the survey ($Z = 1$) and the control group of unselected recipients ($Z = 0$) before the start of the survey. This result is expected, given the random selection process. The last two columns, however, show that respondents to the survey ($D = 1|Z = 1$), i.e., selected cases who will respond to at least one wave of the survey, participate more often in ALMP and in more ALMP than nonrespondents ($D = 0|Z = 1$). These results support the claim that actual participation in the survey, i.e., take-up of the treatment, is selective, which inspired our use of the instrumental variable approach in the first place. We observe similar patterns after waves one, two and three (rows three through eight).

In the last two rows, we see that after participating in the survey, the mean number of days to find a job as well as the percent of people who found a job do not differ much between $Z = 1$ and $Z = 0$. Differences between respondents and nonrespondents, however, are more pronounced. Respondents need less time to find a job and are more likely to find a job. We evaluate whether these differences are due to survey participation in the next section.

3.5.2 Treatment effects

Figure 3.1 shows the results of the intention-to-treat analysis, i.e., the differences in the ALMP outcomes between the group assigned to survey participation and the group assigned to control as shown in Equation 3.1. Results of this analyses are similar to columns one and two of Table 3.3, but the results shown in Figure 3.1 control for the covariates introduced in Section 3.4.5. Before we discuss the ITT estimates, we first check that there are no significant differences between the selected and unselected cases with respect to the two ALMP outcomes measured *prior* to the start of the survey, i.e., we run the falsification test described in Section 3.4.5. The first row in Figure 3.1 shows that the method estimates no significant differences in the number of ALMP (left panel) nor for the ALMP participation indicator (right panel). This finding supports the claim that assignment to treatment or control was in fact random, conditional on X (assumption (i)).

After the first wave (the second row of Figure 3.1), the ITT is positive and significant for both ALMP outcomes, though very small. Unemployment benefit recipients selected for the survey participate in about 0.01 more ALMP than the control group (left panel of Figure 3.1),

Table 3.3: Descriptive statistics of outcome variables

	Outcomes	**Mean (Std. dev.)**			
		Selected	Not selected	Respondents[a]	Non-respondents
		$(Z = 1)$	$(Z = 0)$	$(D = 1 \vert Z = 1)$	$(D = 0 \vert Z = 1)$
Before wave one	Number of ALMP participations	1.71	1.70	1.79	1.68
		(2.05)	(2.03)	(2.07)	(2.04)
	ALMP participation	63%	62%	64%	62%
	N	38,350	38,350	8,728	29,622
After wave one & before wave two	Number of ALMP participations participations	0.35	0.34	0.37	0.35
		(0.71)	(0.70)	(0.71)	(0.71)
	ALMP participation	25%	24%	27%	25%
	N	38,350	38,350	8,647	29,703
After wave two & before wave three	Number of ALMP participations	0.69	0.66	0.71	0.68
		(1.11)	(1.11)	(1.12)	(1.11)
	ALMP participation	37%	36%	39%	37%
	N	38,350	38,350	8,674	29,676
After wave three & before wave four	Number of ALMP participations	1.01	0.98	1.04	1.00
		(1.48)	(1.49)	(1.51)	(1.47)
	ALMP participation	45%	44%	46%	45%
	N	38,350	38,350	8,728	29,622
After wave one & before wave four	Days until new job	1,333.74	1,324.26	1,280.28	1,335.26
		(928.40)	(928.52)	(926.07)	(928.60)
	Found job	54%	55%	58%	54%
	N	38,350	38,350	8,728	29,622

[a] See Table 3.2 for definitions.

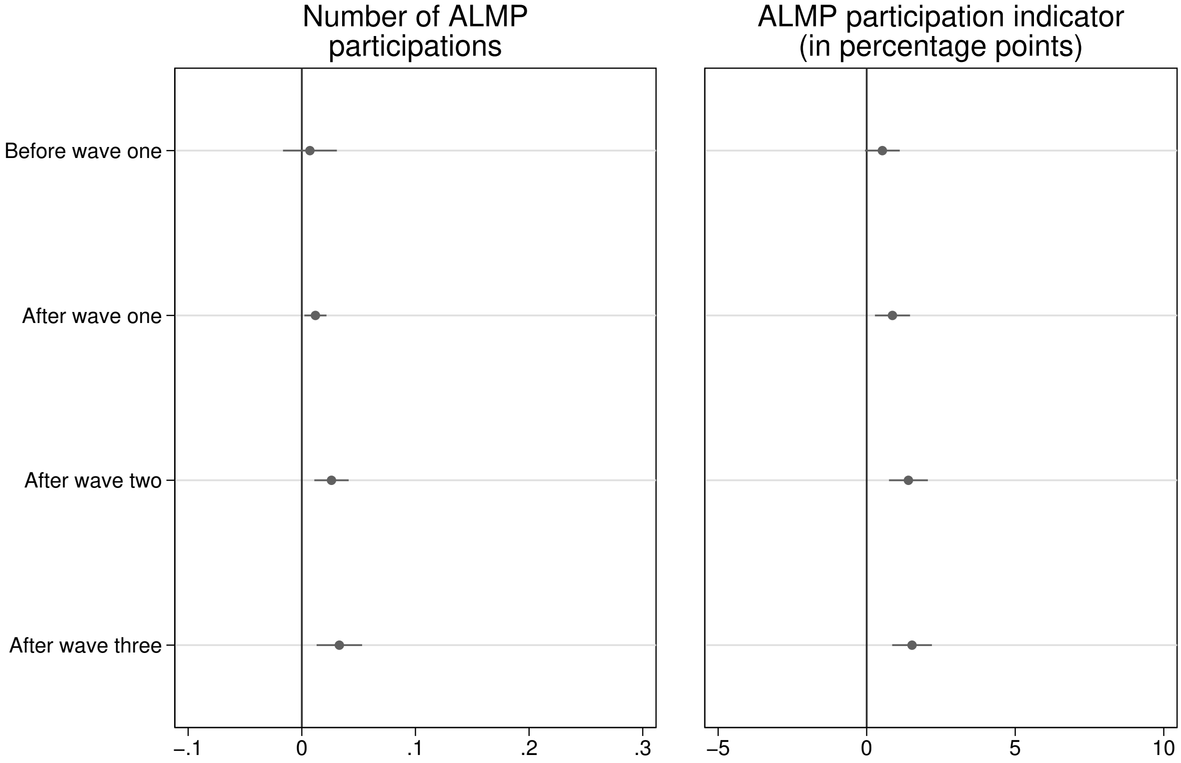

Figure 3.1: Point estimates of panel conditioning effects with 95% confidence intervals. Estimates from Intention-to-treat analysis, controlling for covariates X (see Section 3.4.5). N=76,700.

controlling for X. In addition, take-up of ALMP is about one percentage point larger among the selected cases (right panel of Figure 3.1). Both effects become stronger after waves two and three (rows three and four of Figure 3.1), i.e., when respondents may have been exposed to the survey more than once. These results are evidence that participation in the PASS survey leads to a small increase in ALMP participation, and the effect becomes stronger over the three waves.

Regarding the time to find a job, the ITT is 3.98 days with a standard error of 6.12, controlling for X. The ITT for the indicator of whether a person found a job is -0.17 percentage points with a standard error of 0.33. Both ITTs are insignificant. Thus, ITT results do not seem to support the hypothesis that respondents need less time to find a job or find a job more often than people who were not interviewed.

However, all ITT effects described above underestimate the average causal effect on the treated, due to substantial noncompliance with the treatment assignment status. Using the LARF estimator described in Section 3.4, we can get a clearer picture of the actual effect size of (repeated) survey participation.

Figure 3.2 shows the main results of our study, the LARF estimates from the IV models. Again, we control for the set of covariates X (see Section 3.4.5) to ensure that treatment assignment Z is in fact random (conditional on X) and to reduce some of the variability in the outcome variable and thereby increase the precision of the estimates. Note that once we include covariates in the estimation, the LARF estimator described in Section 3.4 still equals the ATT under one-sided noncompliance. The ATT, however, is then specific to these covariates.

Again, before turning to the outcomes after survey participation, we first look at the take-up of ALMP *before* the start of the survey to assess the model specification with a falsification test. If the model is correctly specified and if the assumptions discussed in Section 3.4 hold, then there should be no significant effect on ALMP participation prior to treatment. The first row of Figure 3.2 reports the results of this analysis. We do not find a significant treatment effect for the number of ALMP (left panel) or for the ALMP participation indicator (right panel). These results, together with the results from the falsification test in Figure 3.1, suggest that

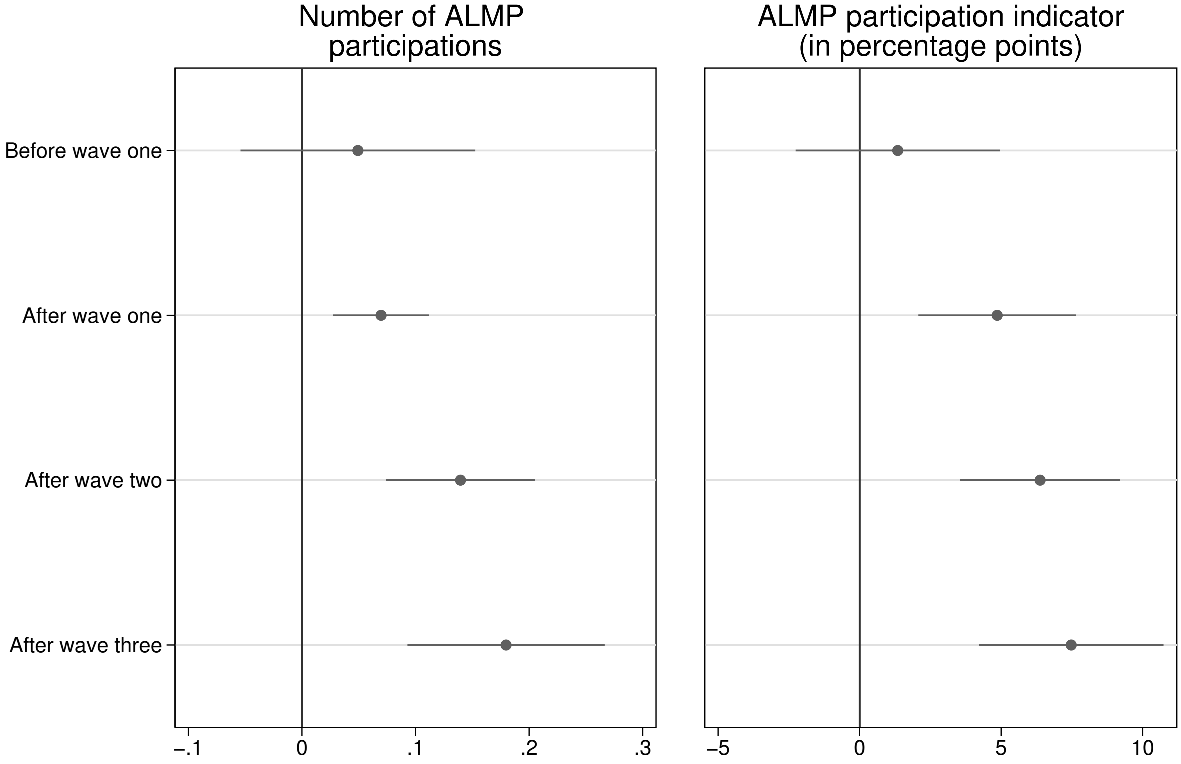

Figure 3.2: Point estimates of panel conditioning effects with 95% confidence intervals. Estimates from instrumental variable analysis using the LARF estimator and controlling for covariates X (see Section 3.4.5). N=76,700.

our IV model is correctly specified, and the assumptions discussed in Section 3.4 are met and thus that we can estimate the causal effect of survey participation on respondents' take-up of ALMP.

Turning to the other rows in Figure 3.2, the effect of participation in the survey on ALMP take-up after the start of the survey, we find a significant positive causal effect of participation in the first wave of the survey on respondents' take-up of ALMP (second row). Respondents participate in about 0.07 more programs (left panel) and participation increases the probability of taking-up an ALMP by about 4.8 percentage points (right panel). After Wave 2, the magnitude of both effects increases to about 0.14 programs and 6.4 percentage points, respectively. Wave 3 leads to a further increase in effect size to about 0.18 and 7.5, respectively. Comparing the LARF estimates (Figure 3.2) to the ITT estimates (Figure 3.1), we see the expected increase in effect size (recall that the ITT underestimates the true effect because it does not adjust for noncompliance). However, we also see that, mainly due to noncompliance among participants offered to participate in the survey, the confidence intervals of the LARF estimates are much wider.

Next, we analyze whether the survey also affects respondents' employment probabilities. Because ALMP are designed to increase employment probabilities for the unemployed, we expect that the increase in ALMP participation leads to a decrease in the time that respondents need to find a job and that respondents find a job more often than people who were not interviewed. Using the same LARF estimator, we do not find that respondents, after participating in up to three waves of the panel, find a job faster (17.22 days with a standard error of 26.80). Similarly, respondents do not find a job more often than people who were not interviewed (-0.73 percentage points with a standard error of 1.43). Thus, we do not find evidence that the increase in ALMP participation results in a significant change in employment probability.

Taken together, these results are strong evidence for the presence of panel conditioning effects in PASS. Survey participation seems to increase the number of programs people participate in and to increase the likelihood to participate in ALMP. Thus, repeated participation in PASS changes respondents' actual labor market behavior, i.e., respondents are more likely to participate in

ALMP, and participate in more programs, than similar persons who are not exposed to the survey. Moreover, these effects seem to increase with the number of survey waves. Contrary to our expectations, survey participation does not affect respondents' employment probabilities. However, taking all the results and the falsification test together, we find strong evidence of changes-in-behavior panel conditioning in the PASS survey.

3.6 Discussion

In this study, we have regarded selection and participation in the first three waves of a large panel survey as a treatment and have used techniques of causal analysis to estimate changes in respondent behavior due to panel conditioning as a treatment effect. The results revealed that respondents of the first three waves are more likely to participate in ALMP, and participate in more programs, than a group of eligible but unselected unemployment benefit recipients. Each additional exposure to the treatment, i.e., each additional wave, intensifies this effect. We do not see an increase of respondents' employment probabilities.

Because many people selected for the survey did not respond, our analysis accounts for non-compliance with the treatment assigned. We relied on an instrumental variable approach and instrumented actual participation in the survey with the random assignment of people to the survey. In addition, we discussed the assumptions necessary to identify the local average response function with this instrument in detail and addressed potential concerns about model misspecification or violated assumptions with a falsification test.

Two theoretical mechanisms offer plausible explanations as to why panel participation can change respondents' behavior. We hypothesized that asking people repeatedly about a specific behavior works as a stimulus, which increases respondents' awareness and motivates them to take up this behavior. Another hypothesis holds that feeling embarrassed about reporting non-normative behavior, people bring their behavior in line with society's norms. In line with these hypotheses, we found that respondents' labor market behavior is changed by (repeated) participation in the survey. However, we cannot distinguish between these two mechanisms.

The true panel conditioning effect may be larger than that shown in Figure 3.2. We likely underestimated the effects of repeated participation in the survey for two reasons. First, we treated all members of a responding household as respondents due to privacy regulations that prevent us from identifying individuals within households. Not all individuals in responding households responded to the survey, however, and therefore did not receive the stimulus. Thus, our treatment group included some untreated cases. Second, the definitions of the treatments after waves two and three also include people who responded in one wave only and therefore received only a 'reduced' stimulus (see Table 3.2).

Our findings may have interesting implications for research on ALMP, labor market policy and labor economics in general. Several studies in the field find that assigning unemployment benefit recipients to an ALMP has a "threat effect" (e.g., Van den Berg and Bergemann 2009; Fitzenberger, Osikominu, and Paul 2010; Graversen and Larsen 2013). That is, unemployed individuals who are assigned to participate in an ALMP increase their job search activity and/or lower their reservation wages to find a job before the program starts, to avoid having to participate in the program. Such an effect may exist because individuals dislike the participation experience or because it reduces their leisure time or time available for job search. Participating in a panel survey that makes ALMP more salient by asking questions about them, however, seems to drive people into program participation. Whether it is preferable to bring people into jobs that pay less than their original reservation wages, or to stimulate them to participate in ALMP through survey participation, however, is a question we cannot answer in this paper, especially because we do not find evidence that the increase in program participation caused by participating in the survey affects respondents' employment probabilities.

Future work should expand our results to other outcomes for which administrative data are available. We would also like to see our results replicated with other data sets and variables. However, we note that in scenarios where external validation records are not available, clear distinction between the two forms of panel conditioning (and elimination of other confounding sources of error) will be difficult. Unfortunately, few studies will have external validation data at hand.

Our results also suggest that the PASS recipient sample is no longer representative of all recipients in Germany (at least in terms of ALMP participation), because participation in the survey over the waves has changed respondents' behavior. As a consequence, inference made from PASS data may be biased if it includes ALMP participation either as a dependent or independent variable. For example, assessments using the PASS data of whether ALMP programs help the unemployed find a new job, an important public policy question, may apply only to PASS respondents and not to the larger recipient population, because the respondents have been changed by the survey.

We note that the participation in federal labor market programs is a rather specific form of (labor market) behavior, and we cannot generalize these findings to other behaviors of interest in panel studies. Yet, with our example, we hope to raise researchers' attention to the fact that repeated participation in panel surveys can change respondents' behavior, a fact that is often not acknowledged by researchers working with panel data.

The possibility that panel data such as PASS is biased due to changes-in-behavior and/or changes-in-reporting panel conditioning has been acknowledged for a long time. However, to date, panel conditioning has been primarily studied by researchers from survey methodology or survey research, and the majority of work has been published in corresponding journals. Yet, as panel data and panel methods have become more popular in recent years with social scientists and economists as tools to uncover causal effects, applied researchers need to be aware that such data can come with new sources of error. Other panel-specific sources of error, such as attrition, have been widely acknowledged by researchers and are addressed, for example, by weighting methods or by introducing refreshment samples. Panel conditioning, by contrast, is often ignored. Our results suggest this strategy is unwise, because panel conditioning can have strong effects on substantively important variables.

3.7 Appendix

3.7.1 Additional analysis Section 3.4.2

The figure below shows associations between covariates X (Y-axis) and the treatment assignment indicator Z. As expected from random assignment of cases to the $Z = 1$ and $Z = 0$, there are only minor differences between the two groups, justifying Assumption (i) (see Section 3.4.2). To increase the precision of our ITT and LARF estimates, we increase all of the covariates shown in the figure (those that differ between $Z = 1$ and $Z = 0$ and those that do not differ between the two groups) in all of our analyses presented in Section 3.5.2.

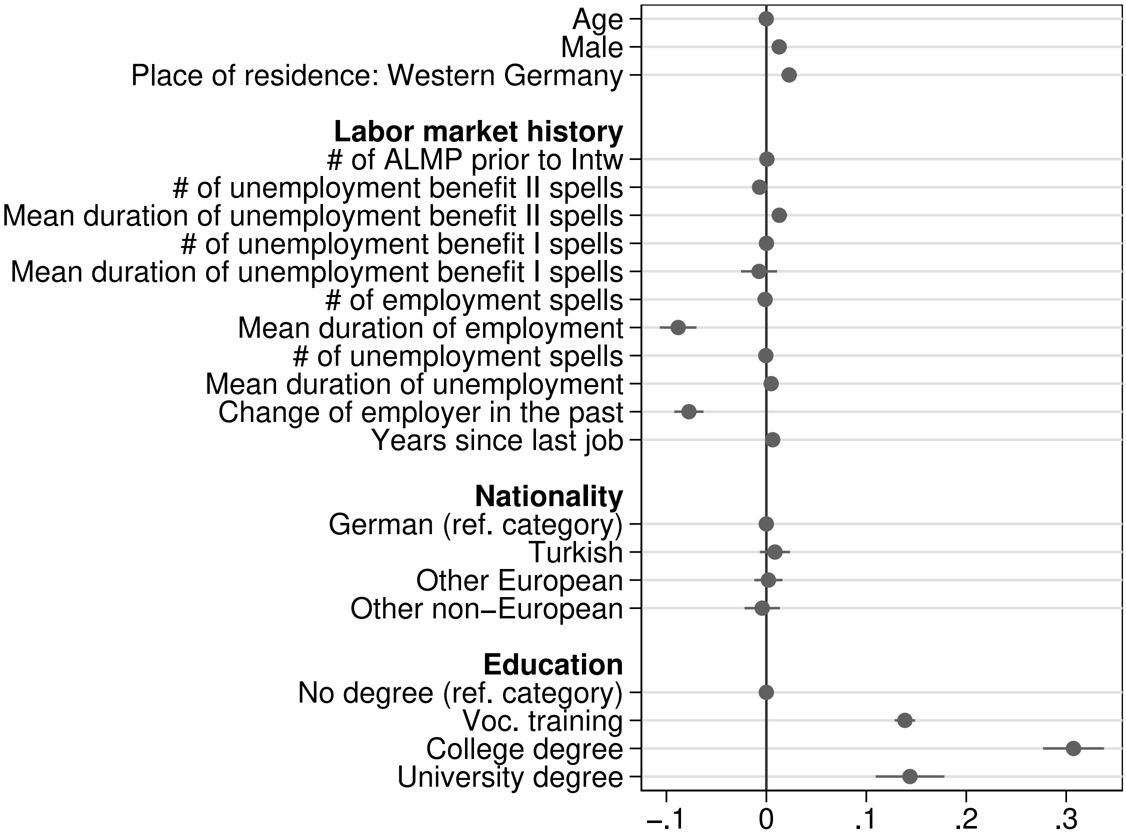

Figure 3.3: Coefficients from OLS regression of covariates X on treatment assignment indicator Z. All covariates measured prior to start of the survey.

Note: Marginal effects from a logistic regression model accounting for the binary nature of Z produce similar results. Unemployment benefit II: longer-term benefits due to unemployment, disability or employment that does not reach a minimum standard of living. Unemployment benefit I: short-term unemployment benefits due to job loss.

3.7.2 Additional analysis Section 3.4.5

Table 3.4: Descriptive statistics of control variables X

	%	Mean	Std. Dev.	Min	Max
Gender					
Male	50.7%				
Female	49.3%				
Place of Residence					
Western Germany	65.9%				
Eastern Germany	34.1%				
Nationality					
German	82.5%				
Turkish	6.0%				
Other European	7.1%				
Non-European	4.4%				
Education					
No degree	28.0%				
Voc. training	69.4%				
College Degree	1.5%				
University Degree	1.1%				
Age		38.11	12.93	15.00	78.38
Labor market history					
# of unemployment benefit II spells		1.48	1.50	0	40
Mean duration of unemployment benefit II		0.42	0.46	0	1.94
# of unemployment benefit I spells		6.96	8.22	0	145
Mean duration of unemployment benefit I spells		0.31	0.26	0	2.66
# of employment spells		11.82	11.33	0	409
Mean duration of employment		0.40	0.25	0	1.00
# of unemployment spells		7.16	6.00	0	81
Mean duration of unemployment		0.61	1.09	0	42.13
Change of employer in the past		0.41	0.27	0	1
Years since last job		1.04	2.99	0.00	31.46

Note: Unemployment benefit II: longer-term benefits due to unemployment, disability or employment that does not reach a minimum standard of living. Unemployment benefit I: short-term unemployment benefits due to job loss.

3.7.3 Official English version of PASS survey questions asking for ALMP participation (Panel 'Labour Market and Social Security' 2006; Panel 'Labour Market and Social Security' 2007; Panel 'Labour Market and Social Security' 2008)

Wave 1

Employment agencies and [fill in type of authority responsible for administration of unemployment benefit II] have various possibilities to support you in finding a vocational training position or a job. Now we would like to learn about your experiences. Have you since January 2005 participated in a program financed or promoted by the job center ("Arbeitsamt") or [name of authority responsible for unemployment benefit II], which was to improve your prospects for finding a job or a vocational training position like application training measures for instance, or in a program that offered you a job opportunity like a one-euro job for example? Please also think of programs that were of short duration, maybe of a few days only.

Wave 2

Have you participated in at least one of the following programs, measures or courses, which was promoted or financed by the job center ("Arbeitsamt") or [fill in type of authority responsible for administration of unemployment benefit II] since January 2006? Please also consider programs which were only of short duration.

Wave 3

Have you January 2007 participated in at least one of the following programs, measures or courses which was financed or promoted by the job center ("Arbeitsamt") or employment agency? Please also consider programs which were only of short duration.

References

Abadie, A. (2003). "Semiparametric Instrumental Variable Estimation of Treatment Response Models". In: *Journal of Econometrics* 113, pp. 231–263.

An, W. and X. Wang (2016). "LARF: Instrumental Variable Estimation of Causal Effects through Local Average Response Functions". In: *Journal of Statistical Software* 71.1, pp. 1–13.

Angrist, J.D., G.W. Imbens, and D.B. Rubin (1996). "Identification of Causal Effects Using Instrumental Variables". In: *Journal of the American Statistical Association* 91.434, pp. 444–472.

Angrist, J.D. and J.-S. Pischke (2009). *Mostly Harmless Econometrics. An Empiricist's Companion.* Princeton, NY: Princeton University Press.

Athey, Susan and Guido W. Imbens (2017). "The State of Applied Econometrics: Causality and Policy Evaluation". In: *Journal of Economic Perspectives* 31.2, pp. 3–32.

Berg, Marco, Ralph Cramer, Christian Dickmann, Daniel Gebhardt, Reiner Gilberg, Birgit Jesske, Karen Marwinski, Claudia Wenzig, and Martin Wetzel (2010). "Codebook and Documentation of the Panel Study 'Labour Market and Social Security' (PASS) - Datenreport Wave 3". In: *FDZ Datenreport 06/2010.*

Borle, Sharad, Uptal M. Dholakia, Siddharth S. Singh, and Robert A. Westbrook (2007). "The Impact of Survey Participation on Subsequent Customer Behavior: An Empirical Investigation". In: *Marketing Science* 26.5, pp. 711–726.

Bundesagentur für Arbeit (2017). *Bedarfsgemeinschaften und deren Mitglieder - Deutschland mit Ländern und Kreisen (unrevidiert) - Juli 2006.* Available at `https://statistik.arbeitsagentur.de/Statistikdaten/Detail/200607/iiia7/gs-asu-sgbii-rev/gs-asu-sgbii-rev-d-0-xls.xls`, last accessed Jan 23, 2018.

Cantor, David (2010). "A Review and Summary of Studies on Panel Conditioning". In: *Handbook of Longitudinal Research. Design, Measurement and Analysis.* Ed. by S. Menard. Amsterdam: Elsevier, pp. 123–138.

Chandon, Pierre, Vicki G. Morwitz, and Werner J. Reinartz (2005). "Do Intentions Really Predict Behavior? Self-Generated Validity Effects in Survey Research". In: *Journal of Marketing* 69.2, pp. 1–14.

Christoph, Bernhard, Gerrit Müller, Daniel Gebhardt, Claudia Wenzig, Mark Trappmann, Juliane Achatz, Anita Tisch, and Christine Gayer (2008). "Codebook and Documentation of the Panel Study 'Labour Market and Social Security' (PASS)". In: *FDZ Datenreport 05/2008* vol. 1: Introduction and overview, wave 1 (2006/2007).

Clausen, Aage R. (1969). "Response Validity: Vote Report". In: *The Public Opinion Quarterly* 32, pp. 588–606.

Clinton, Joshua D. (2001). *Panel Bias from Attrition and Conditioning: A Case Study of the Knowledge Networks Panel.* Tech. rep. Stanford, CA: Department of Political Science Technical Report, Stanford University.

Crépon, B. and G.J. Van den Berg (2016). "Active Labor Market Policies". In: *Annual Review of Economics* 8, pp. 521–546.

Crossley, Thomas, Jochem de Bresser, Liam Delaney, and Joachim Winter (2017). "Can Survey Participation Alter Household Saving Behavior?" In: *Economic Journal* 127, pp. 2332–2357.

Das, Marcel, Vera Toepoel, and Arthur van Soest (2011). "Nonparametric Tests of Panel Conditioning and Attrition Bias in Panel Surveys". In: *Sociological Methods & Research* 40.1, pp. 32–56.

Dennis, J M. (2001). "Are Internet Panels Creating Professional Respondents? A Study of Panel Effects". In: *Marketing Research* 13.2, pp. 34–39.

Dholakia, Uptal M. (2010). "A Critical Review of Question-Behavior Effect Research". In: *Review of Marketing Research* 7, pp. 147–199.

Duan, Naihua, Margarita Alegria, Glorisa Canino, Thomas McGuire, and David Takeuchi (2007). "Survey Conditioning in Self-Reported Mental Health Service Use: Randomized Comparison of Alternative Instrument Formats". In: *Health Research and Educational Trust* 42.2, pp. 890–907.

Fitzenberger, Bernd, Aderonke Osikominu, and Marie Paul (2010). "The Heterogeneous Effects of Training Incidence and Duration on Labor Market Transitions". In: *IZA Discussion Paper*

Series No. 5269. Available at `ftp://repec.iza.org/RePEc/Discussionpaper/dp5269.pdf`, last accessed Jan 23, 2018.

Gebhardt, Daniel, Gerrit Müller, Arne Bethmann, Mark Trappmann, Bernhard Christoph, Christine Gayer, Bettina Müller, Anita Tisch, Bettina Siflinger, Hans Kiesl, Bernadette Huyer-May, Juliane Achatz, Claudia Wenzig, Helmut Rudolph, Tobias Graf, and Anika Biedermann (2009). "Codebook and Documentation of the Panel Study 'Labour Market and Social Security' (PASS)". In: *FDZ Datenreport 06/2009* Volume I: Introduction and overview. Wave 2 (2007/2008).

Granberg, Donald and Soren Holmberg (1992). "The Hawthorne Effect in Election Studies: The Impact of Survey Participation on Voting". In: *British Journal of Political Science* 22.2, pp. 240–247.

Graversen, Brian Krogh and Brian Larsen (2013). "Is there a Threat Effect of Mandatory Activation Programmes for the Long-term Unemployed?" In: *Empirical Economics* 44, pp. 1031–1051.

Groves, Robert M., Robert M. Cialdini, and Mick P. Couper (1992). "Understanding the Decision to Participate in a Survey". In: *Public Opinion Quarterly* 56.4, pp. 475–495.

Groves, Robert M., Eleanor Singer, and Amy Corning (2000). "Leverage-Saliency Theory of Survey Participation: Description and an Illustration". In: *Public Opinion Quarterly* 64.3, pp. 299–308.

Halpern-Manners, Andrew and John R. Warren (2012). "Panel Conditioning in Longitudinal Studies: Evidence From Labor Force Items in the Current Population Survey". In: *Demography* 49.4, pp. 1499–1519.

Halpern-Manners, Andrew, John R. Warren, and Florencia Torche (2017). "Panel Conditioning in the General Social Survey". In: *Sociological Methods & Research* 46.1, pp. 103–124.

Imbens, G.W. and J.D. Angrist (1994). "Identification and Estimation of Local Average Treatment Effects". In: *Econometrica* 62.2, pp. 467–475.

Institute for Employment Research (2013). *Integrated Employment Biographies. V11.00.00.* Nuremberg.

Jacobebbinghaus, Peter and Stefan Seth (2007). "The German Integrated Employment Biographies Sample IEBS". In: *Schmollers Jahrbuch. Zeitschrift für Wirtschafts- und Sozialwissenschaften* 127.2, pp. 335–342.

Jacobi, Lena and Jochen Kluve (2007). "Before and After the Hartz Reforms: The Performance of Active Labour Market Policy in Germany". In: *Journal for Labour Market Research* 40.1, pp. 45–64.

Köhler, Markus and Ulrich Thomsen (2009). "Data Integration and Consolidation of Administrative Data from Various Sources. The Case of Germans' Employment Histories". In: *Historical Social Research* 34.3, pp. 215–229.

Kraut, Robert E. and John B. McConahay (1973). "How Being Interviewed Affects Voting: An Experiment". In: *The Public Opinion Quarterly* 37.3, pp. 398–406.

Kreuter, Frauke, Gerrit Müller, and Mark Trappmann (2010). "Nonresponse and Measurement Error in Employment Research. Making Use of Administrative Data". In: *Public Opinion Quarterly* 74.5, pp. 880–906.

Lazarsfeld, Paul F. (1940). "'Panel Studies'". In: *The Public Opinion Quarterly* 4, pp. 122–128.

Lepkowski, J. and M. Couper (2002). "Nonresponse in the Second Wave of Longitudinal Household Surveys". In: *Survey Nonresponse*. Ed. by R. Groves, J. Eltinge D. Dillman, and R. Little. New York: Wiley, pp. 259–271.

Morwitz, Vicki G., Eric Johnson, and David Schmittlein (1993). "Does Measuring Intent Change Behavior?" In: *The Journal of Consumer Research* 20.1, pp. 46–61.

Nancarrow, Clive and Trixie Cartwright (2007). "Online Access Panels and Tracking Research: The Conditioning Issue". In: *International Journal of Market Research* 49.5, pp. 573–594.

Panel 'Labour Market and Social Security' (2006). *Questionnaire Wave One. (Available at http://doku.iab.de/fdz/pass/Questionnaires_English.zip., last accessed Jan 23, 2018)*. URL: %7Bhttp://doku.iab.de/fdz/pass/Questionnaires_English.zip.%7D.

— (2007). *Questionnaire Wave Two. (Available at http://doku.iab.de/fdz/pass/Questionnaires_English_W2.zip., last accessed Jan 23, 2018)*. URL: %7Bhttp://doku.iab.de/fdz/pass/Questionnaires_English_W2.zip.%7D.

Panel 'Labour Market and Social Security' (2008). *Questionnaire Wave Three. (Available at **http://doku.iab.de/fdz/pass/Questionnaires_English_W3.zip.**, last accessed Jan 23, 2018).* URL: `%7Bhttp://doku.iab.de/fdz/pass/Questionnaires_English_W3.zip.%7D`.

Pennell, Steven G. and James N. Lepkowski (1992). "Panel Conditioning Effects in the Survey of Income and Program Participation". In: *Proceedings of the American Statistical Association, Survey Research Methods Section*, pp. 566–571.

R Core Team (2017). *A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing.

Rizzo, L., G. Kalton, and M. Brick (1996). "A Comparison of Some Weighting Adjustment Methods for Panel Nonresponse". In: *Survey Methodology* 22, pp. 43–53.

Rubin, Donald B. (1974). "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies". In: *Journal of Educational Psychology* 66.5, pp. 688–701.

— (1977). "Assignment to Treatment Group on the Basis of a Covariate". In: *The Journal of Educational Statistics* 2, pp. 1–26.

— (1978). "Bayesian Inference for Causal Effects: The Role of Randomization". In: *Annals of Statistics* 6, pp. 34–58.

Rudolph, H. and M. Trappmann (2007). "Design und Stichprobe des Panels "Arbeitsmarkt und Soziale Sicherung" (PASS)". In: *Neue Daten für die Sozialstaatsforschung. Zur Konzeption der IAB-Panelerhebung "Arbeitsmarkt und Soziale Sicherung" IAB Forschungsbericht 12 / 2007.* Ed. by ed. M. Promberger. Nürnberg, pp. 60–101.

Schnell, R. (2007). "Alternative Verfahren zur Stichprobengewinnung für ein Haushaltspanelsurvey mit Schwerpunkt im Niedrigeinkommens- und Transferleistungsbezug". In: *Neue Daten für die Sozialstaatsforschung. Zur Konzeption der IAB-Panelerhebung "Arbeitsmarkt und Soziale Sicherung" IAB Forschungsbericht 12 / 2007.* Ed. by ed. M. Promberger. Nürnberg, pp. 33–59.

Shack-Marquez, Janice (1986). "Effects of Repeated Interviewing on Estimation of Labor-Force Status". In: *Journal of Economic and Social Measurement* 14.4, pp. 379–398.

Smith, Jennifer K., Alan S. Gerber, and Anton Orlich (2003). “Self-Prophecy Effects and Voter Turnout: An Experimental Replication”. In: *Political Psychology* 24.3, pp. 593–604.

Spangenberg, Eric and Carl Obermiller (1996). “To Cheat or Not to Cheat: Reducing Cheating by Requesting Self-Prophecy”. In: *Marketing Education Review* 6.3, pp. 95–103.

Stock, J.H. and M. Yogo (2005). “Testing for Weak Instruments in Linear IV Regression”. In: *Identification and Inference for Econometric Models. Essays in Honor of Thomas Rothenberg.* Ed. by D.W.K. Andrews and J.H. Stock. New York: Cambridge University Press, pp. 80–108.

Sturgis, Patrick, Nick Allum, and Ian Brunton-Smith (2009). “Attitudes Over Time: The Psychology of Panel Conditioning”. In: *Methodology of Longitudinal Surveys.* Ed. by P. Lynn. New York: Wiley, pp. 113–126.

Toepoel, Vera, Marcel Das, and Arthur van Soest (2009). “Relating Question Type to Panel Conditioning: Comparing Trained and Fresh Respondents”. In: *Survey Research Methods* 3.2, pp. 73–80.

Trappmann, Mark, Jonas Beste, Arne Bethmann, and Gerrit Müller (2013). “The PASS Panel Survey After Six Waves”. In: *Journal for Labour Market Research* 46.4, pp. 275–281.

Traugott, Michael W. and John P. Katosh (1979). “Response Validity in Surveys of Voting Behavior”. In: *Public Opinion Quarterly* 43.3, pp. 359–377.

Van den Berg, Gerard J. and Annette H. Bergemann (2009). “The Effect of Labor Market Programs on Not-Yet Treated Unemployed Individuals”. In: *Journal of the European Economic Association* 7.2-3, pp. 606–616.

Van der Zouwen, Johannes and Theo Van Tilburg (2001). “Reactivity in Panel Studies and its Consequences for Testing Causal Hypotheses”. In: *Sociological Methods & Research* 30.1, pp. 35–56.

Warren, John R. and Andrew Halpern-Manners (2012). “Panel Conditioning in Longitudinal Social Science Surveys”. In: *Sociological Methods & Research* 41.4, pp. 491–534.

Waterton, Jennifer and Denise Lievesley (1989). “Evidence of Conditioning Effects in the British Social Attitudes Panel”. In: *Panel Surveys.* Ed. by D. Kasprzyk, G. J. Duncan, G. Kalton, and M. P. Singh. New York: Wiley, pp. 319–339.

Williams, Patti, Lauren G. Block, and Gavan J. Fitzsimons (2006). "Simply Asking Questions about Health Behaviors Increases both Healthy and Unhealthy Behaviors". In: *Social Influence* 1.2, pp. 117–127.

Williams, William H. and Colin L. Mallows (1970). "Systematic Biases in Panel Surveys Due to Differential Nonresponse". In: *Journal of the American Statistical Association* 65.331, pp. 1338–1349.

Yalch, Richard F. (1976). "Pre-Election Interview Effects on Voter Turnout". In: *Public Opinion Quarterly* 40.3, pp. 331–336.

Yan, Ting and Stephanie Eckman (2012). "Panel Conditioning: Change in True Value versus Change in Self-Report". In: *Proceedings of the American Statistical Association, Survey Research Methods Section*, pp. 4726–4736.

Zabel, J.E. (1998). "An Analysis of Attrition in the Panel Study of Income Dynamics and the Survey of Income and Program Participation with an Application to a Model of Labor Market Behavior". In: *Journal of Human Resources* 33.2, pp. 479–506.

Zwane, Alix Peterson, Jonathan Zinman, Eric van Dusen, William Pariente, Clair Null, Edward Miguel, Michael Kremer, Dean S. Karlan, Richard Hornbeck, Xavier Giné, Esther Duflo, Florencia Devoto, Bruno Crepon, and Abhijit Banerjee (2011). "Being Surveyed Can Change Later Behavior and Related Parameter Estimates". In: *Proceedings of the National Academy of Sciences* 108.5, pp. 1821–1826.

Chapter 4

Motivated Misreporting in Web Panels

4.1 Abstract

Previous studies of reporting to filter questions have shown that respondents learn to say “No” to filter questions to shorten the interview, a phenomenon called motivated misreporting. Similar learning effects have been observed in panel surveys: Respondents seem to recall the structure of a survey from earlier waves and, in subsequent waves, give responses that shorten the interview. Hence, concerns arise that misreporting to filter questions worsens over time in a panel study. We conducted an experiment using filter questions in two consecutive waves of a monthly online panel to study how misreporting to filter questions changes over time. While we replicate previous findings on the filter question format effect, we do not find any support for the hypothesis that responses to filter questions worsen over time. Our findings add to the literature on data quality in web panels, panel conditioning and motivated misreporting.

4.2 Background

Motivated misreporting refers to the phenomenon whereby respondents deliberately give inaccurate survey responses to reduce the burden of the survey. Motivated misreporting has been shown either within a single survey *or* over two or more waves of a panel survey, but we are not aware of any study that has analyzed motivated misreporting within *and* between waves of a panel study. We test whether motivated misreporting worsens over two waves of a panel survey. To begin with, we review relevant findings and theoretical explanations in the literature.

Many surveys use filter questions to determine respondent eligibility for follow-up questions. Their purpose is to reduce burden for respondents by asking only relevant questions. Two common formats are used for these filter questions. In the *interleafed* format, respondents are asked a filter question, and the follow-ups, if triggered, follow immediately. In the *grouped* format, respondents are first asked all filters before answering the follow-ups that apply. Several studies have shown that respondents trigger fewer follow-ups in the interleafed format than in the grouped format (Kessler et al. 1998; Duan et al. 2007; Kreuter et al. 2011; Eckman et al. 2014). In the interleafed format, respondents learn how the survey is structured and begin to say "No" to the filter questions to avoid the follow-up questions. Comparisons with administrative records have shown that the grouped format collects more accurate reports (Eckman et al. 2014). Consistent with respondents' learning, the difference in reporting between the two filter question formats is larger for items asked later in a survey (Kreuter et al. 2011). Duan et al. (2007) call such misreporting due to learning the structure of the interview *survey conditioning*, because respondents' reporting behavior is conditioned by survey participation.

A similar mechanism underlies one form of panel conditioning. Respondents may learn in an earlier wave how the questionnaire is structured and use this information to misreport. They then become *worse reporters* over time in a panel survey (Bailar 1989; Waterton and Lievesley 1989; Cantor 2010; Yan and Eckman 2012). Other forms of panel conditioning are also possible, such as respondents becoming better reporters (Kroh, Winter, and Schupp 2016) or panel conditioning resulting in changes in respondents' attitudes (Warren and Halpern-Manners 2012) or behavior (Bach and Eckman 2017). However, given the evidence from survey conditioning, i.e., respondents' tendency to become worse reporters due to learning the structure of the questionnaire, we focus on the worse respondents mechanism in this study.[1]

[1] An extensive discussion of theoretical mechanisms underlying panel conditioning, their outcomes and a review of relevant studies is provided by Warren and Halpern-Manners (2012).

Halpern-Manners and Warren (2012) find evidence for the worse reporters hypothesis in the Current Population Survey. Some respondents interviewed for the second time overreport employment. The authors suspect that respondents remember from the first round that reporting unemployment triggers extensive follow-up questions and thus misreport unemployment in the second wave to skip follow-up questions. However, given the topic of unemployment, social desirability may also influence respondents' reports. Additional support for the worse reporters hypothesis has been found in reports of home alteration and repair jobs (Neter and Waksberg 1964), functional limitations of elderly people (Mathiowetz and Lair 1994) and every day personal hygiene product use (Nancarrow and Cartwright 2007). Schonlau and Toepoel (2015) find support for the worse reporters hypothesis in the LISS panel, the panel study that also provides the data for our analysis. They show that straightlining, i.e., the tendency to give the same responses to a series of questions with identical answer choices, increases with respondents' panel experience. On the other hand, using different studies, Cohen and Burt (1985) and Struminskaya (2016) find no evidence that panel respondents become worse reporters over time. Regarding the likelihood of the occurrence of panel conditioning, Van Landeghem (2012) reports that panel conditioning becomes more likely the higher the number of exposures to the same survey. Moreover, Halpern-Manners, Warren, and Torche (2014) report that panel conditioning is more likely the shorter the interval between waves.

Although the studies cited above have analyzed panel and survey conditioning separately, the literature falls short of testing these two effects jointly. Given the consistent findings on survey conditioning and the evidence from previous studies of panel conditioning, a careful test of both phenomena in a joint context is needed, the more so, as both forms of conditioning share the same theoretical mechanism, a desire by respondents to reduce the burden of the survey. In this paper, we use a two by two design where half of the respondents in wave one receive the questions in the interleafed format and half of the respondents in the grouped format. In wave two, half of the respondents are assigned to the same format as in wave one and the other half switch to the other format.

In this study, we test three hypotheses. We first replicate findings on misreporting to filter questions, i.e., survey conditioning. In line with previous research, we expect that the percent of filter questions triggered will be greater in a grouped format than in an interleafed format in wave one. We call this first hypothesis the *survey conditioning hypothesis.* In addition to replicating survey conditioning, we explore whether survey conditioning depends on the education of a respondent. To date, there is no evidence regarding an interaction between survey conditioning and education or ability.

Second, we aim to determine whether motivated misreporting persists over waves. Kreuter et al. (2011) show that misreporting in an interleafed format is worse in later items in a survey and conclude that respondents learn how the survey is structured. When respondents are presented with the same questions in wave two, they already know the structure and can misreport from the very first filter question. Thus, we hypothesize that respondents become worse respondents, i.e., the percent of filters triggered will be smaller in the second wave among respondents interviewed in the same format in both waves. We call this the *panel conditioning hypothesis.* Respondents in the interleafed format can learn in wave one and underreport in wave two. Respondents in the grouped format have no opportunity to learn in wave one, but they may nevertheless recall the structure in wave two. Consequently, we expect a decrease in filter questions triggered in wave two among respondents interviewed in the same format in both waves.

Third, we study how panel conditioning affects misreporting when respondents are interviewed in different formats in each wave, i.e., when the format changes from grouped to interleafed (and vice versa) over time. We refer to the effects of changing the format over time as the *changed format panel conditioning hypothesis.* Recognizing response patterns in this context and disentangling panel and survey conditioning (i.e., changes in reporting due to changing the format) is harder, as both panel conditioning and the format may influence responses at the same time, possibly even in different directions. Switching from grouped to interleafed, misreporting should increase over time, due to both survey conditioning and panel conditioning. However, switching from interleafed to grouped, the effect is somewhat more complex. We expect that respondents recall the questionnaire from the first wave and thus, due to panel conditioning, misreporting should increase: this is the worse reporters mechanism. At the same time, however, the grouped format usually collects more reports. Thus, we would expect a decrease in misreporting after switching the format. Consequently, we cannot predict whether misreporting will increase or decrease over time for this group. Theory does not predict whether gains in data quality from changing to the better grouped format are outweighed by losses due to panel conditioning. Yet, this question is especially important for panel surveys using the interleafed format that wish to change to the grouped format to reduce motivated misreporting.[2] Table 4.1 summarizes the second and third hypotheses and the expected effects on the level of misreporting.

In addition to testing these three hypotheses, we explore whether both survey conditioning and panel conditioning depend on respondents' cognitive ability. Evidence from studies on response order effects, non-differentiation, and satisficing show that respondents with low education are most susceptible to such effects (e.g., Krosnick and Alwin 1987; Narayan and Krosnick 1996; Holbrook et al. 2007). Similarly, Binswanger, Schunk, and Toepoel (2013) find a panel conditioning effect in difficult attitudinal questions only among low educated respondents. In our case, survey conditioning may on the one hand increase with cognitive ability, because respondents with higher cognitive abilities may be faster in discovering repetitive patterns of the questionnaire. On the other hand, respondents with higher cognitive abilities might instead show less misreporting in the interleafed format as they may be more aware of the importance of accurate survey reports for scientific purposes. The same arguments may hold for panel conditioning: Respondents with higher cognitive ability may be better at recalling having answered the same questionnaire one month ago. However, their better understanding of the importance of accurate survey reports for research may again counteract.

4.3 The study

The data we use for our analysis of survey and panel conditioning comes from the Dutch LISS panel, a longstanding Internet panel based on a probability sample. Sample members complete

[2] Other panel surveys might consider a switch from the grouped to the interleafed format, for reasons we do not discuss here, for brevity. See Kreuter, Eckman, and Tourangeau (forthcoming) for a discussion of the advantages of the interleafed format.

Table 4.1: Panel conditioning hypotheses

Filter question format		**Hypothesis**	**Effect on**
Wave one	Wave two		**misreporting**
Interleafed	Interleafed	*Panel conditioning*	Increase[a]
Grouped	Grouped	*Panel conditioning*	Increase[a]
Grouped	Interleafed	*Changed format panel cond.*	Increase[a]
Interleafed	Grouped	*Changed format panel cond.*	Unclear

[a] *Corresponds to a decrease in the number of filters triggered.*

online questionnaires of about 15 to 30 minutes on a monthly basis (Scherpenzeel and Das 2010).

In 2012, we put several filter question experiments in two consecutive waves of the LISS panel. In the first wave (April 2012), LISS participants (n=3,330) were randomly assigned to either the interleafed or grouped filter question format. In the second wave (May 2012), all panel members were again randomly assigned to one of the two formats. The resulting design has four cells (two formats by two waves), as shown in Table 4.5.[3]

In the interleafed format, respondents were asked 13 filter questions in random order. Two follow-up questions were asked immediately after each filter, if applicable. In the grouped format, respondents were asked all 13 filter questions (in random order) before answering any applicable follow-ups. Apart from the filter questions and a handful of questions asking for respondents' experience with the questionnaire, no other questions were asked in this module.[4] Given the small number of questions, the median response time was a little more than three minutes in both waves and 95% of all respondents answered the questionnaire in less than eight minutes.

All filter questions asked about purchases of items such as groceries, clothes or tickets for movies during the last month (see the Appendix for the original questions). We chose these questions to ensure that most respondents triggered at least a few follow-up questions in each wave. Also, we chose items that should not be influenced by circumstances such as seasonality that would

[3] There was a third format included in the original experiment. However, we do not use these cases. Results concerning that format are reported in Kreuter, Eckman, and Tourangeau (forthcoming).

[4] While no other questions were asked in this module, respondents are usually invited to answer more than one module per month. We do not have any information on other modules our respondents answered in these months.

Table 4.2: Response rates by filter question format

Filter question format in wave one		**Wave one** n	**Wave one** Response Rate[a]	**Wave two** n[b]	**Wave two** Response rate[c]
Interleafed	Respondents	1,303	78.9%	1,080	81.8%
	Nonrespondents	347		242	
Grouped	Respondents	1,337	81.0%	1,082	82.0%
	Nonrespondents	313		238	
Total number of respondents		2,640	79.3%	2,162	81.9%
|Z-statistic|[d]			1.48		0.10
p-value			0.139		0.919

[a] *AAPOR RR1.*

[b] *Among wave one respondents.*

[c] *AAPOR RR1, conditional on wave one response.*

[d] *H_0: % triggered interleafed = % triggered grouped*

lead to a real change of purchasing behavior between the two waves of the experiment. Thus, we expect that any differences in reporting between the two waves are, on average, caused by a change in reporting and not by a change in behavior.

Nonresponse and Attrition

About 79% of the LISS panel members selected for the study participated in the first wave of the experiment (AAPOR RR1). Conditional on wave one response, the participation rate in wave two is 82% (third column in Table 4.2). For the analysis sample of our research questions, however, we discard those who responded only in wave one (n=478) or only in wave two (n=326). Panel attrition can easily be mistaken for panel conditioning, and disentangling the two effects is one of the major challenges when analyzing panel conditioning (Bach and Eckman 2017). Using only two wave respondents allows us to eliminate any confounding effects of attrition (Warren and Halpern-Manners 2012). Two additional respondents are excluded as they broke off the survey, one in wave one and one in wave two. The resulting analysis sample consists of 2,162 people (third column in Table 4.2).

Before we report the results of our analysis, however, we make sure that sub-setting the analysis sample to two-wave respondents does not confound the experimental manipulations regarding the filter question formats and thus our estimates of survey and panel conditioning. To do so, we check whether nonresponse in wave one and attrition differ by the two formats and waves. Regarding nonresponse in wave one, response rates do not differ between the two formats (second column in Table 4.2), and we do not find any substantial or significant differences in socio-demographic variables between the formats (results not shown). Moreover, nonresponse over time, i.e., attrition, does not seem to be related to socio-demographic variables: we find only one substantial and significant difference between those who respond to both waves and those who participate in the first wave only. Attriters are younger, but do not differ from two-wave respondents in gender, education, marital status, household composition or in their response behavior in previous waves of the LISS panel (results not shown). Importantly, neither the filter question format in wave one, the number of filters triggered in wave one, nor their interaction is predictive of attrition (Table 4.3). Wave two response rates do not differ between the two wave one formats (fourth column in Table 4.2).

Since *all* LISS participants were randomly assigned to the two filter question formats in wave two, we also test whether there are any important differences between the two-wave-respondents and all respondents of wave two, i.e., whether dropping some respondents from wave two interferes with the randomization to the filter question formats in wave two. After sub-setting the sample to two-wave-respondents, we do not find any significant differences in socio-demographic variables or the number of filters triggered in wave one between the two formats in wave two (results not shown). Therefore, we are confident that nonresponse and attrition patterns do not differ among the subgroups. Importantly, attrition is not influenced by the experimental conditions or the filter question experience in wave one and thus should not bias the results of our analysis. Although the sample restrictions limit the representativeness of the data (external validity), they do not jeopardize our interpretation of the experimental results (internal validity).

In the next two sections, we present and discuss the results of our analysis regarding our three hypotheses, i.e., whether respondents learn to misreport within a single wave of the survey and whether misreporting persists over two waves.

Table 4.3: Predictors of attrition

	Coefficient
Grouped (ref.: interleafed)	0.031 (0.037)
No. of filters triggered in wave one	0.002 (0.005)
Grouped*No. of filters triggered in wave one	-0.002 (0.007)
N	2,640[a]

Note: Linear probability model: Attrition (Yes/No). Standard errors in parentheses. Constant not shown.
[a]*All wave one respondents.*

4.4 Results

As expected by the survey conditioning hypothesis, the percent of filter questions triggered in wave one is significantly smaller in the interleafed format than in the grouped format (36% vs. 43%, i.e., 4.7 vs. 5.6 filters triggered, Table 4.4). Survey conditioning is taking place in wave one as we expected. When we also include the position of a filter question (i.e., whether a filter was asked in the first, second or third third of the section) in the analysis, we find that the probability of triggering a filter in the interleafed format decreases significantly with the position of a filter in the questionnaire (results not shown). This finding further supports our first hypothesis, that respondents learn to misreport in the interleafed format.

Regarding the second hypothesis, panel conditioning, we expect that the percent of filters triggered should decrease over time due to learning when filter questions are asked in the same format in each wave (rows one and two of Table 4.1). Row one of Table 4.5 shows that respondents interviewed in the interleafed format in *both* waves show no significant change in reporting across time. The same result holds for respondents interviewed in the grouped format in *both* waves. Thus, we do not find any support for the panel conditioning hypothesis: Respondents do not trigger fewer filters when interviewed for the second time in the same format.

Table 4.4: Percent of filter questions triggered in wave one, by format

Format	**Wave one** % triggered
Interleafed	36.3 (0.005)
Grouped	43.1 (0.006)
Total	39.7 (0.004)
|Z-statistic|[a]	84.83
p-value	<0.001
n	2,162

Note: Standard errors (in parentheses) clustered at the respondent level.

[a] *H_0: % triggered interleafed = % triggered grouped*

Next, we turn to our third hypothesis and those respondents who changed formats from wave one to wave two (rows three and four of Table 4.1). We expect that respondents switched from grouped to interleafed underreport in wave two due to panel conditioning (see Table 4.1). Row three of Table 4.5 shows, as expected, that these respondents trigger significantly fewer filters in wave two than in wave one and thus increase misreporting. Switching from interleafed to grouped, however, leads to a significant increase in filters triggered and thus less misreporting (row four). Both effects have about the same absolute size. Moreover, the effect size is approximately equal to the difference between the interleafed format and the grouped format in wave one. For this reason, we interpret the changes over time seen in the bottom rows of Table 4.5 as driven by the format change and not by panel conditioning. Thus, we find no support for the third hypotheses that panel conditioning influences misreporting when the filter question format changes between waves.

In both panel conditioning hypotheses, we hypothesize that respondents learn from participation in the first wave of the experiment how follow-up questions can be avoided. However, a few respondents (n=48) did not trigger a single filter question in wave one. Therefore, these respondents could not learn. To assess whether this confounds the results shown in Table 4.5, we rerun the analyses without these 48 respondents. Results without these respondents do not differ substantially from the results shown in Table 4.5 (results not shown).

Regarding the influence of cognitive ability, we find that survey conditioning does not depend

Table 4.5: Percent of filter questions triggered, by experimental condition

Format		**Percent of filters triggered**		n	|Z-statistic|[a]	p-
Wave one	Wave two	Wave one	Wave two			value
Interleafed	Interleafed	36.1 (0.007)	36.5 (0.007)	571	0.86	0.353
Grouped	Grouped	43.3 (0.007)	43.4 (0.007)	565	0.00	0.948
Grouped	Interleafed	42.8 (0.008)	35.0 (0.007)	517	116.34	<0.001
Interleafed	Grouped	36.4 (0.007)	43.3 (0.008)	509	107.87	<0.001

Note: Standard errors (in parentheses) clustered at the respondent level.
[a]*H_0: % triggered wave one = % triggered wave two*

on cognitive ability, measured by respondents' highest educational degree achieved. In other words, the difference between filters triggered in the interleafed format and filters triggered in the grouped format does not vary with respondents' education (results not shown). Similarly, we do not find a panel conditioning effect when replicating the results of Table 4.5 in subgroups defined by respondents' highest educational degree. We interpret these findings as evidence that survey conditioning and panel conditioning do not depend on respondents' cognitive ability.

To sum up the results presented here, we find strong evidence for survey conditioning. Fewer filters are triggered in the interleafed format than in the grouped format, additional evidence that respondents in the interleafed format learn how the survey is structured and use that information to reduce the burden of the survey by misreporting. However, there is no evidence that respondents remember the survey structure over time in a panel survey context and misreport more in the second wave and thus no support for the worse respondents mechanism.

4.5 Discussion

Survey researchers are increasingly aware that asking filter questions in different formats can affect responses and measurement error. Studies on misreporting in cross-sectional surveys have shown that respondents asked in an interleafed format trigger fewer follow-ups than respondents interviewed in a grouped format. The common explanation for such survey conditioning is that respondents learn how a survey is structured and use this information to speed through the interview and reduce the burden of the survey. Similar concerns have been raised in the context of panel surveys: If respondents remember the structure of a survey from prior waves, they might recall that giving certain answers can reduce the length of the survey.

We hypothesized that first, respondents surveyed in the interleafed filter question format would misreport more than respondents surveyed in the grouped format. We found in both waves of our experiment that respondents in the interleafed format trigger fewer filters than respondents in the grouped format. Our interpretation of this finding is that misreporting is more pronounced in the interleafed format: Studies that compared responses to filter questions with validation records (see Section 4.2) showed that the grouped format produces more correct responses. Thus, we are confident in our interpretation of the grouped format as more accurate than the interleafed format. These results are as expected from previous studies.

In our second and third hypotheses, we suspected that we would see more misreporting in a second wave of a panel study. If respondents remember the structure of a questionnaire from prior participation, then misreporting in a second wave should increase: respondents would recognize the filter questions and trigger fewer filters in a second wave. In this study, the two waves were separated by only four weeks, which should make panel conditioning more likely (Warren and Halpern-Manners 2012; Halpern-Manners, Warren, and Torche 2014). However, we did not find support for either of the two panel conditioning hypotheses. Respondents interviewed in the same format in both waves do not misreport more in the second wave (the panel conditioning hypothesis). We did see differences in triggering rates for those respondents whose formats changed across waves (the changed format panel conditioning hypothesis). However, these changes in the levels of misreporting are likely only due to the format itself and not due to the worse respondents (or better respondents) mechanism. Interestingly, neither survey conditioning nor panel conditioning seem to vary with respondents' cognitive ability, as measured by respondents' highest educational degree.

To ensure that our findings are not distorted by attrition, we considered only those respondents who participated in both waves of the survey. Furthermore, we chose a battery of filter questions that should not lead to changes-in-behavior panel conditioning. One might argue that the absence of panel conditioning is due to LISS respondents being very good or very motivated respondents who are not susceptible to motivated misreporting. Yet, the study by Schonlau and Toepoel (2015) regarding straightlining (see Section 4.2) and the clear signs of misreporting in the interleafed format in wave one (see Section 4.4) show that LISS respondents are as good (or bad) as any other respondents. Even so, we did not find a panel conditioning effect in our filter questions.

This lack of a panel conditioning effect is somewhat surprising. While respondents show a strong tendency to remember the structure of a questionnaire in a single survey, they do not seem to recall this information when the same questions are repeated four weeks later. Even respondents in the interleafed format in wave one, many of whom misreported, no longer show any signs of misreporting when interviewed for a second time in a grouped format. The most likely explanation for this finding is the resetting effect reported by Kreuter et al. (2011). They find that respondents' misreporting is reset when a new section within a questionnaire starts. A similar logic might apply to the panel survey context: If misreporting resets when a new section starts, then the same should hold for the start of a new wave in a panel survey.

The degree of panel conditioning seems to depend on the topic and the burden of the questions asked in a survey (Warren and Halpern-Manners 2012). The questions we asked in the LISS panel were fairly easy to answer, dealt with a possibly boring topic ("What did you buy last month?") and respondents needed only a few minutes to answer them (see Section 4.3). In addition, LISS respondents were exposed to our questionnaire only twice and have answered many other, potentially more interesting, novel or burdensome LISS modules in the past. Responding to our filter questions module may not have been distinct from the general experience of participating in the LISS panel, and thus respondents may not have remembered the questions (Tourangeau, Rips, and Rasinksi 2000, chapter 3). Questions dealing with more burdensome or sensitive topics may produce different results, potentially even resulting in respondents becoming better reporters over time (compare the studies cited in Section 4.2). Moreover, results may differ when respondents are exposed to the survey more than twice (see Section 4.4). Thus, more research exploring the influence of topic and survey burden as well as the level of exposure to the survey on motivated misreporting and panel conditioning needs to be done, a point also made by Eckman et al. (2014).

Our finding, however, that motivated misreporting does not increase over waves of a panel survey is good news for survey practitioners. Although misreporting can pose a serious problem, repeated participation in the same survey does not make the problem worse. Importantly, for panel surveys that used the interleafed format in the past, our work suggests that changing to a grouped format is a good idea, as it can help increase data quality by eliminating survey conditioning without introducing panel conditioning.

4.6 Appendix

Filter questions

- In the past month, have you purchased coffee for consumption at home?
- In the past month, have you purchased beer or wine for consumption at home?
- In the past month, have you purchased tobacco?
- In the past month, have you purchased children's clothing or shoes?
- In the past month, have you purchased clothing or shoes for yourself?
- In the past month, have you purchased chocolate?
- In the past month, have you purchased medication?
- In the past month, have you purchased flowers?
- In the past month, have you purchased pet supplies?
- In the past month, have you purchased movies on DVD or VHS?
- In the past month, have you purchased music on CD or as MP3s (or other
- In the past month, have you purchased a ticket for a concert, theater
- In the past month, have you purchased any cleaning supplies for your home?

Follow-up questions

For each yes answer to the above filter questions:

Thinking about your most recent purchase of (fill: item)

- How much did it cost?
  - (Open ended response in Euros)

  - a. Don’t know
  - b. Refused

- For whom was it purchased?
  - a. self
  - b. another household member
  - c. someone else
  - d. Don’t know
  - e. Refused

References

Bach, Ruben L. and Stephanie Eckman (2017). “Does Participating in a Panel Survey Change Respondents’ Labor Market Behavior?” In: *IAB Discussion Paper* 15/2017. Available at `http://doku.iab.de/discussionpapers/2017/dp1517.pdf`, last accessed Jan 23, 2018.

Bailar, Barbara A. (1989). “Information Needs, Surveys, and Measurement Errors”. In: *Panel Surveys.* Ed. by D. Kasprzyk, G. J. Duncan, G. Kalton, and M. P. Singh. New York: Wiley, pp. 1–24.

Binswanger, Johannes, Daniel Schunk, and Vera Toepoel (2013). “Panel Conditioning in Difficult Attitudinal Questions”. In: *Public Opinion Quarterly* 77.3, pp. 783–797.

Cantor, David (2010). “A Review and Summary of Studies on Panel Conditioning”. In: *Handbook of Longitudinal Research. Design, Measurement and Analysis.* Ed. by S. Menard. Amsterdam: Elsevier, pp. 123–138.

Cohen, Steven B. and Vicki L. Burt (1985). “Data Collection Frequency Effect in the National Medical Care Expenditure Survey”. In: *Journal of Economic & Social Measurement* 13.2, pp. 125–151.

Duan, Naihua, Margarita Alegria, Glorisa Canino, Thomas McGuire, and David Takeuchi (2007). “Survey Conditioning in Self-Reported Mental Health Service Use: Randomized Comparison of Alternative Instrument Formats”. In: *Health Research and Educational Trust* 42.2, pp. 890–907.

Eckman, Stephanie, Frauke Kreuter, Antje Kirchner, Annette Jäckle, Roger Tourangeau, and Stanley Presser (2014). “Assessing the Mechanisms of Misreporting to Filter Questions in Surveys”. In: *Public Opinion Quarterly* 78.3, pp. 721–733.

Halpern-Manners, Andrew and John R. Warren (2012). “Panel Conditioning in Longitudinal Studies: Evidence From Labor Force Items in the Current Population Survey”. In: *Demography* 49.4, pp. 1499–1519.

Halpern-Manners, Andrew, John R. Warren, and Florencia Torche (2014). “Panel Conditioning in a Longitudinal Study of Illicit Behaviors”. In: *Public Opinion Quarterly* 78.3, pp. 565–590.

Holbrook, A. L., Jon A. Krosnick, David Moore, and Roger Tourangeau (2007). “Response Order Effects in Dichotomous Categorical Questions Presented Orally: The Impact of Question and Respondent Attributes”. In: *Public Opinion Quarterly* 71.3, pp. 1–25.

Kessler, Ronald C., Hans-Ulrich Wittchen, Jamie M. Abelson, Katherine McGonagle, Norbert Schwarz, Kenneth S. Kendler, Bärbel Knäuper, and Shanyang Zhao (1998). “Methodological Studies of the Composite International Diagnostic Interview (CIDI) in the US National Comorbidity Survey (NCS)”. In: *International Journal of Methods in Psychiatric Research* 7.1, pp. 33–55.

Kreuter, Frauke, Stephanie Eckman, and Roger Tourangeau (forthcoming). “Salience of Burden and Its Effects on Response Behavior to Skip Questions. Experimental Results from Telephone and Web-Surveys”. In: *Advances in Questionnaire Design, Development, Evaluation and Testing.* Ed. by P. Beatty, D. Collins, L. Kaye, J. Padilla, G. Willis, and A. Wilmot. Hoboken, NJ: Wiley.

Kreuter, Frauke, Susan McCulloch, Stanley Presser, and Roger Tourangeau (2011). “The Effects of Asking Filter Questions in Interleafed Versus Grouped Format”. In: *Sociological Methods & Research* 40.1, pp. 88–104.

Kroh, Martin, Florin Winter, and Juergen Schupp (2016). “Using Person-Fit Measures to Assess the Impact of Panel Conditioning on Reliability”. In: *Public Opinion Quarterly* 80.4, pp. 914–942.

Krosnick, Jon A. and D. F. Alwin (1987). “An Evaluation of a Cognitive Theory of Response-Order Effects in Survey Measurement”. In: *Public Opinion Quarterly* 51.2, pp. 201–219.

Mathiowetz, Nancy A. and Tamra J. Lair (1994). “Getting Better? Change or Error in the Measurement of Functional Limitations”. In: *Journal of Economic and Social Measurement* 20.3, pp. 237–262.

Nancarrow, Clive and Trixie Cartwright (2007). “Online Access Panels and Tracking Research: The Conditioning Issue”. In: *International Journal of Market Research* 49.5, pp. 573–594.

Narayan, Sowmya and Jon A. Krosnick (1996). “Education Moderates Some Response Effects in Attitude Measurement”. In: *Public Opinion Quarterly* 60.1, pp. 58–88.

Neter, John and Joseph Waksberg (1964). “Conditioning Effects from Repeated Household Interviews”. In: *Journal of Marketing* 28.2, pp. 51–56.

Scherpenzeel, Annette and Marcel Das (2010). ““True” Longitudinal and Probability-Based Internet Panels: Evidence From the Netherlands”. In: *Social and Behavioral Research and the Internet: Advances in Applied Methods and Research Strategies.* Ed. by M. Das, P. Ester, and L. Kaczmirek. Boca Raton: Taylor & Francis, pp. 77–104.

Schonlau, Matthias and Vera Toepoel (2015). “Straightlining in Web Survey Panels Over Time”. In: *Survey Research Methods* 9.2, pp. 125–137.

Struminskaya, Bella (2016). “Respondent Conditioning in Online Panel Surveys. Results of Two Field Experiments”. In: *Social Science Computer Review* 34.1, pp. 95–115.

Tourangeau, Roger, Lance J. Rips, and Kenneth Rasinksi (2000). *The Psychology of Survey Response.* Cambridge: Cambridge University Press.

Van Landeghem, Bert (2012). “Panel Conditioning and Subjective Well-Being: Evidence from International Panel Data and Repeated Cross-Sections”. In: *SOEPPaper* No. 484.

Warren, John R. and Andrew Halpern-Manners (2012). “Panel Conditioning in Longitudinal Social Science Surveys”. In: *Sociological Methods & Research* 41.4, pp. 491–534.

Waterton, Jennifer and Denise Lievesley (1989). “Evidence of Conditioning Effects in the British Social Attitudes Panel”. In: *Panel Surveys.* Ed. by D. Kasprzyk, G. J. Duncan, G. Kalton, and M. P. Singh. New York: Wiley, pp. 319–339.

Yan, Ting and Stephanie Eckman (2012). “Panel Conditioning: Change in True Value versus Change in Self-Report”. In: *Proceedings of the American Statistical Association, Survey Research Methods Section*, pp. 4726–4736.

Chapter 5

Motivated Misreporting Among Reluctant Respondents

5.1 Abstract

Many surveys aim to achieve high response rates to keep bias due to nonresponse low. However, research has shown that the relationship between the nonresponse rate and nonresponse bias is small. In fact, high response rates may lead to measurement error, if respondents with low response propensities provide survey responses of low quality. In this paper, we explore the relationship between response propensity and measurement error, specifically motivated misreporting, the tendency to give inaccurate answers to speed through an interview. Using data from five surveys conducted in several countries and modes, we analyze whether motivated misreporting is worse among those respondents who were the least likely to respond to the survey. Contrary to the prediction of our theoretical model, we do not find clear evidence that reluctant respondents are more likely to misreport.

5.2 Background

Many surveys aim to achieve high response rates to keep bias due to nonresponse low, but increasing the response rate by bringing in reluctant respondents may lead to measurement error. That is, respondents who are the least likely to become respondents may provide survey responses of low quality when they do respond (Curtin, Presser, and Singer 2000; Curtin, Presser, and Singer 2005; Groves, Couper, and Dipko 2004; Groves 2006; Groves and Peytcheva 2008; Keeter et al. 2000; Merkle and Edelman 2002; Tourangeau, Groves, and Redline 2010; Peytchev, Peytcheva, and Groves 2010; Olson 2013). Thus, researchers who use extraordinary measures to increase the response rate may in fact increase total error (Biemer 2001; Groves 2006).

To study the relationship between respondents' reluctance and measurement error, we must operationalize both reluctance and measurement error. To measure respondents' reluctance, we estimate response propensities, i.e., the probability of each person who was selected for a survey to respond to the survey. Respondents with the lowest response propensities are reluctant respondents. We operationalize measurement error through motivated misreporting, a phenomenon whereby respondents deliberately give inaccurate or false responses to reduce the burden of the survey. This response behavior is often observed in questions used to determine respondent eligibility for follow-up questions. Asking such questions in certain formats allows respondents to learn how follow-up questions can be avoided by giving inaccurate or false answers, thus introducing measurement error. The motive behind this motivated misreporting is respondents' desire to reduce the burden of the survey (Tourangeau, Kreuter, and Eckman 2015). Respondents who have a low propensity to respond to the survey at all may be more interested than other respondents in reducing the burden of the survey when they do respond. Thus, reluctant respondents should show more motivated misreporting, supporting the hypothesis that response propensity affects measurement error. We elaborate on these operational definitions and the hypothesis in more detail in the next section.

To study this hypothesis empirically, we use five surveys that were conducted in three countries (the Netherlands, the U.S. and Germany) and modes (Online, CAPI and CATI). Each contained

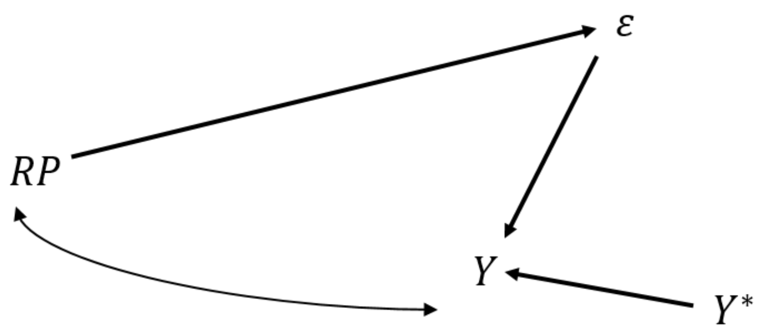

Figure 5.1: Nonresponse-Measurement Error model explaining a relationship between response propensity (RP) and measurement error (ϵ) in a reported survey variable Y. Y^* is the true value of the outcome (figure adapted from Groves (2006)).

experimental manipulations of questions prone to motivated misreporting: filter questions and looping questions. These experimental manipulations allow us to study the connection between response propensity and measurement error. Before we review the data in more detail, we present the theoretical reasoning underlying the hypothesis that nonresponse influences measurement error.

5.3 A Nonresponse-Measurement Error Model

The idea that reluctant respondents may be worse reporters builds on the nonresponse-measurement error model developed by Groves (2006), shown in Figure 5.1. This model suggests a nexus between response propensity and measurement error. The reported value of Y equals the true value, Y^*, plus an error term ϵ, i.e., for individual i, $Y_i = Y_i^* + \epsilon_i$. The magnitude of the error for case i, ϵ_i, is determined by the response propensity of the case RP_i. Applying the model to our study, low response propensities cause high levels of measurement error, while high response propensities cause lower levels of measurement error.

This model does not specify how exactly response propensity influences the error term. Many possible mechanisms exist. For example, lack of interest in the survey topic can cause a case to have a low response propensity (Martin 1994; Groves, Couper, and Dipko 2004) and may also explain why low interest respondents who *do* participate in the survey put less effort into answering survey questions carefully and truthfully. Other motives such as a general reluctance to help out (Tourangeau, Groves, and Redline 2010) or a lack of motivation and cooperativeness (Cannell and Fowler 1963; Bollinger and David 2001) may also reduce RP and cause ϵ. Thus, there may be some characteristics Z that explain both RP and ϵ and induce the relationship between the two shown in Figure 5.1. These external causes are excluded from the model. Nevertheless, we can use this model to test our hypothesis about the relationship between response propensity and motivated misreporting. Before we present the analytical strategy, we review previous findings in the literature on the connection between nonresponse and measurement error in the next section.

5.4 Previous findings

Empirical studies of a connection between (non)response propensity and measurement error have focused on several aspects of both response propensity and measurement error. Cannell and Fowler (1963), for example, assess the impact of nonresponse on reporting errors in self-reported hospital stays. Comparing self-reports with administrative hospital records, they find that respondents who needed extensive follow-up tend to misreport both the number of hospital stays and their duration. However, it is unclear if the higher level of measurement error among late respondents is caused by response propensity or simply the result of the increased recall period for late respondents (Fricker and Tourangeau 2010). Kreuter, Müller, and Trappmann (2010), also validating survey reports against administrative records, find that measurement error among respondents recruited with increased levels of follow-up offsets the reduction in nonresponse bias gained by including them in terms of total survey error. Other studies show that late respondents (Willimack et al. 1995) and converted refusers (Triplett et al. 1996) have higher item nonresponse rates. Little evidence, however, is found by Keeter et al. (2000) regarding the effects of more rigorous recruiting strategies compared to standard recruiting strategies on item nonresponse. Studies using response propensity scores find that including low response propensity cases results only in a weak increase in measurement error that is offset by gains in reduction of nonresponse bias (Olson 2006). Peytchev, Peytcheva, and Groves (2010) show that low response propensity cases underreport abortion experiences, show more misreporting errors in voting behavior (Tourangeau, Groves, and Redline 2010) and higher item nonresponse rates (Fricker and Tourangeau 2010). Low response propensity cases, however, do not show more acquiescence, extreme responses or non-differentiation (Yan, Tourangeau, and Arens 2004) or provide answers of worse data quality to questions asking for well-being (Hox, de Leeuw, and Chang 2012). To sum up, the majority of previous research examining the influence of nonresponse on measurement error finds that response propensity does affect measurement error. However, there are some studies where the connection is small or even nonexistent. The different operationalizations of respondents' reluctance and measurement error may explain some of the variation of the findings.

In this study, we use the estimated response propensity scores as the operationalization of respondents' reluctance. The advantage of this approach is that the estimated response propensity score is a comprehensive measure of different aspects of response propensity. That is, depending on the predictors entered into the response propensity model, this score may capture a variety of aspects of response propensity such as the extent of follow-up needed (Cannell and Fowler 1963), how early or late a case responded to the survey (Willimack et al. 1995) and interest in the survey (Martin 1994). Due to this encompassing nature of the response propensity score, we prefer the response propensity score to the more specific measures of reluctance used in other studies. Next, we briefly present findings regarding motivated misreporting, our measure of measurement error.

Three question types, filter questions, looping questions, and screener questions, are prone to motivated misreporting. These questions are typically used to determine respondents' eligibility for follow-up questions. Filter questions, for example used in the National Survey on Drug Use and Health, the U.S. Consumer Expenditure Survey or the National Crime Victimization Survey, are usually asked either in the interleafed or in the grouped format. Respondents in the *interleafed* format are asked a filter question with the follow-ups, if triggered, right away. In the *grouped* format, however, respondents are first asked all filter questions before answering the follow-ups that apply (see Table 5.1 for an example). Looping questions, for example part of the Panel Survey on Income Dynamics, the Health and Retirement Survey or the Panel Survey Labor Market and Social Security, are used to collect data about several similar events, e.g., doctor visits or periods of unemployment. For every event reported, respondents are asked follow-up questions. These looping questions can also be asked in two formats. In the *go-again* format, respondents are asked about the most recent event, for example, and then about details of this event right away. Only then are respondents asked whether there is another event to be reported. In the *how-many* format, however, respondents are first asked about the number of events and then follow-up questions about each event reported (see Table 5.2 for an example). The third type of questions, screener questions, are used to determine if anyone in a household is eligible for the survey. Tourangeau, Kreuter, and Eckman (2012) provide details on the formate of these questions and their effect on motivated misreporting. However, we do not review these questions in more detail because they are not included in our analysis.

Table 5.1: Example of Interleafed vs. Grouped format (filter questions)

Interleafed version	**Grouped version**
In the past 3 months, have you purchased a coat?	In the past 3 months, have you purchased a coat?
Please briefly describe the most recent coat you purchased.	In the past 3 months, have you purchased a shirt? In the past 3 months, have you purchased pants?
For whom was it purchased?	In the past 3 months, have you purchased a suit?
In what month did you purchase it?	In the past 3 months, have you purchased a dress?
How much did it cost?	FOR EACH YES
In the past 3 months, have you purchased a shirt?	Please briefly describe the most recent [item] you purchased.
Please briefly describe the most recent shirt you purchased.	For whom was it purchased?
[...]	In what month did you purchase it?
In the past 3 months, have you purchased a suit?	How much did it cost?
[...]	
[...]	

Note: Table adapted from Kreuter et al. (2011)

Table 5.2: Example of Go-again vs. How-many (looping questions)

Go-again version	**How-many version**
Please think about the first event X that you ever experienced.	How many events X have you experienced in your life?
[Follow-up questions about event]	*[Number of events]*
Did you ever experience another event X?	Thinking about the first event X ...
[Follow-up questions about event]	*[Follow-up questions about event]*
Did you ever experience another event X?	Thinking about the second event X ...
[Follow-up questions about event]	*[Follow-up questions about event]*
...	...

Note: Table adapted from Eckman and Kreuter (forthcoming).

Comparisons between the interleafed and the grouped format in filter questions have shown that respondents trigger fewer follow-ups in the interleafed format than in the grouped format (Kessler et al. 1998; Duan et al. 2007; Kreuter et al. 2011; Eckman et al. 2014). This motivated misreporting is not possible in the grouped format because there is no chance for respondents to learn how the questions work. Similarly, comparisons between the go-again and the how-many format in looping questions have shown that respondents report fewer events in the go-again format than in the how-many format (Eckman and Kreuter, forthcoming). Similar to the interleafed format in filter questions, respondents learn in the go-again format how the survey is structured and begin to report fewer events to avoid the follow-up questions. Comparisons with administrative validation data have shown that misreporting is in fact more pronounced in the interleafed format and the go-again format, respectively (Eckman et al., 2014; Eckman and Kreuter, forthcoming). Thus, differences between the two formats in each question type reflect real differences in measurement error, with the grouped and how-many formats collecting more correct responses to the initial filter or looping questions.

Tests of different mechanisms that could explain the observed format effects have shown that motivated misreporting arises from respondents' desire to reduce the burden of the survey (Eckman et al., 2014). That desire may also affect the response propensity score. For example, respondents who want to reduce burden may be unlikely to respond to the survey at all. The desire to reduce the burden of the survey would therefore be a mechanism that affects both motivated misreporting, ϵ in Figure 5.1, and the response propensity score, RP. The level of measurement error associated with a reported survey outcome would therefore be related to the response propensity score, inducing the relationship shown in Figure 5.1.

Given this theoretical model and the evidence from previous studies, we analyze the connection between response propensity and motivated misreporting. That is, we study whether motivated misreporting is more pronounced among reluctant respondents using reports from several surveys briefly described in the next section.

5.5 Data

Data for our analysis come from five surveys, conducted in different countries and modes. We briefly present key characteristics of each survey below and in Table 5.3. The original questions from each survey are shown in the Appendix.

The first survey was conducted as part of the Dutch LISS panel, a longstanding probability-based internet panel. Sample members complete online questionnaires of about 15 to 30 minutes on a monthly basis (Scherpenzeel 2011). In 2012, we put several filter question experiments in two consecutive waves of the LISS panel using the same questionnaire in both waves. In the first wave (April), LISS participants (n=5,513) were randomly assigned to either the interleafed or grouped filter question format. In the second wave (May), participants (n=5,668) were again randomly assigned to one of the two formats.[1] Respondents in both formats were asked 13 filter questions with two follow-up questions for each filter answered with "Yes". All filter questions asked about purchases of items such as groceries, clothes or movie tickets during the last month. About 68 percent (n=3,767) of the LISS panel members selected for the study participated in the first wave of the experiment (AAPOR RR1, AAPOR, 2016) and about 64 percent (n=3,601) participated in wave two. Participation in the second wave was open to all panel members, irrespective of participation in wave one. Since there is no evidence that measurement error increases from wave one to wave two due to panel conditioning (Bach and Eckman forthcoming), we treat each wave separately in our analysis. We refer to the first wave of this survey as LISS-1 and to the second wave as LISS-2. Results regarding motivated misreporting in both LISS-1 and LISS-2 are reported in Bach and Eckman (forthcoming).

The second survey, the Survey on Free Time (SOFT), was a CAPI survey conducted in 2013 in the U.S.. 1,120 households were selected from the U.S. Postal Service's Delivery Sequence File using a three-stage sampling design.[2] The response rate (AAPOR RR1) was about 27

[1] There was another format included in the original experiment. Since we do not use these cases, there are small differences between the number of participants in wave one and wave two. Results concerning the third format are reported in Kreuter, Eckman, and Tourangeau (forthcoming).

[2] Primary sampling units (PSU) were compromised of individual cities or urban areas. Secondary sampling units (SSU) used ZIP codes or ZIP code fragments within sampled PSUs. Participants were then sampled within SSUs.

percent (n=304). Respondents were randomly assigned to answer 16 filter questions in the interleafed format or in the grouped format. Filter questions asked about interest in sports, clothing purchases and watching television, followed by up to six follow-up questions.

The third survey, "Employment and Purchase Behavior in Germany" (EPBG), was a CATI survey conducted in Germany in 2011. 12,400 adults were selected from German administrative labor market records. The response rate was about 19 percent (AAPOR RR1) and we use 1,200 out of 2,400 completed cases in this analysis.[3] Respondents of the EPBG survey were asked 18 filter questions either in the interleafed or in the grouped format, covering clothing purchases, employment history and income sources. Four follow-up questions were asked for each filter, if applicable. We refer to this survey as EPBG-FQ. Results regarding motivated misreporting in this survey are reported in Eckman et al. (2014).

The fourth survey, also called "Employment and Purchase Behavior in Germany" (EPBG), was a web survey conducted in Germany in 2012 (Eckman and Kreuter, forthcoming). While EPBG-FQ contained filter questions, this survey contained looping questions instead. For this reason, we refer to this survey as EPBG-LQ. The sample of 11,836 adults was selected from the same German administrative labor market records and 1,068 cases (about nine percent, AAPOR RR1) responded to the looping questions of the survey.[4] The survey contained two sections of looping questions, one asked about employers and the other one about places the respondent had lived. The order of the questions was randomized and respondents were asked these questions either in the how-many or the go-again format. Results regarding motivated misreporting in this survey are reported in Eckman and Kreuter (forthcoming).

See Table 5.3 for an overview of the five datasets. Four of these datasets (LISS-1, LISS-2, SOFT, EPBG-FQ) contain filter questions and one (EPBG-LQ) contains looping questions. In each, respondents were randomly assigned to the different question formats (interleafed or grouped for the filter questions and go-again or how-many for the looping questions).

[3] An additional 1,200 respondents completed the survey but were assigned to experimental conditions not used in this paper. Results concerning the other experiments are reported in Sakshaug, Tutz, and Kreuter (2013).

[4] The survey contained two additional experiments: An incentive experiment (Felderer, Kreuter, and Winter 2013) and a consent experiment (Sakshaug and Kreuter 2014). We do not expect these manipulations to affect our results.

Table 5.3: Summary of five datasets from four surveys

	LISS-1	**LISS-2**	**SOFT**	**EPBG-FQ**	**EPBG-LQ**
Country	NL	NL	U.S.	Germany	Germany
Mode	Web	Web	CAPI	CATI	Web
Data collection	April 2012	May 2012	April-June 2013	Aug-Oct 2011	Feb-April 2012
Motivated misreporting	FQ	FQ	FQ	FQ	LQ
Question formats	I/G	I/G	I/G	I/G	GA/HM
n *respondents*	3,767	3,601	304	1,200	1,068
Response rate[a]	68%	64%	27%	19%	9%

FQ: Filter questions. LQ: Looping questions.

I: Interleafed. G: Grouped. GA: Go-again. HM: How-many.

[a] *AAPOR RR1 2016.*

To test our hypothesis, we need an estimate of the response propensity score to identify reluctant and non-reluctant respondents as well as a measure of measurement error (that is, the extent of motivated misreporting).

5.6 Estimation of response propensity and motivated misreporting

The idea of the response propensity builds on the seminal work of Rosenbaum and Rubin (1983) on propensity scores. Originally introduced in the field of evaluation studies, the propensity score denotes the conditional probability that a unit (e.g., a person) receives a treatment, given observable attributes of the unit. Similarly, the *response* propensity is the conditional probability that a person responds to a survey or not, given the person's attributes (Bethlehem, Cobben, and Schouten 2011, chapter 11). This score, RP_i, varies between zero and one and is a latent variable. Although we cannot observe it, we can observe the corresponding response indicator, R_i, which allows us to estimate response propensity scores.

Logistic regression is the most common technique for estimating response propensities (Bethlehem, Cobben, and Schouten 2011, chapter 11). The dependent variable in these models is the binary response indicator, R_i, indicating whether a unit responded to a survey or not. All variables known or assumed to influence whether a unit is a respondent to a survey are included in the model as covariates, often in various functional forms (e.g., linear, quadratic, or interacted with other predictors). Predictions from this model then form the response propensity scores. In recent years, however, nonparametric prediction algorithms from machine learning methods have been introduced in the response propensity score literature (McCaffrey, Ridgeway, and Morral 2004; Phipps and Toth 2012; Buskirk and Kolenikov 2010). The major advantage of these methods is that the researcher does not need to determine the (correct) functional form of the predictor variables in the propensity score model, including the decision about which variables should enter the model at all. Rather, machine learning or data mining methods automatically select covariates predictive of the response variable based on the available data. In addition, machine learning methods can deal with many covariates even if the sample size is small. Simulation studies have shown that these methods often outperform standard approaches such as logistic regression in the estimation of (response) propensity scores (e.g., Lee, Lessler, and Stuart 2010; Buskirk and Kolenikov 2010).

5.6.1 Predictors of response

We use boosted regression modeling to estimate the response propensity of each selected case, and we fit a separate model for each of the five datasets. The dependent variable in each model is a binary variable indicating whether a case responded to the survey or not. The independent variables in these models are all variables available for both respondents and nonrespondents in the dataset.[5] As shown in Table 5.3, the five surveys were selected from different sampling frames and conducted in different modes (i.e., online, CAPI, and CATI). Therefore, the information available to build the response propensity model differs between the surveys. Below we describe the variables available for response propensity score prediction in each survey. The appendix shows the complete list of covariates in each model.

LISS-1 and LISS-2 were both conducted as part of the longstanding LISS online panel. Panel members have responded to several other waves of the panel before taking part in our two surveys, and thus the amount of information available for both respondents and nonrespondents

[5] Covariates may contain missing values as long as there are non-missing values observed among both respondents and nonrespondents.

from previous waves is large. All in all, we included 116 covariates in the response propensity model for LISS-1. These covariates cover socio-demographic information (e.g., age, gender, education, employment), a series of attitudinal questions and the response behavior in previous waves, information regarding household composition, as well as paradata from the initial recruitment interview for the panel. The response propensity model for LISS-2 includes the same information as LISS-1 plus information that was collected as part of LISS-1, i.e., whether a person responded in LISS-1, the filter question format and the number of filters triggered in LISS-1.

Compared to LISS, the amount of information available about both respondents and nonrespondents in SOFT is much smaller. The SOFT response propensity model includes covariates derived from paradata that were collected during the CAPI interviews, such as the date and time of the first contact attempt and whether a person ever refused the interview, as well as covariates derived from the sampling design, for example primary and secondary sampling unit identifiers.

The samples of EPBG-FQ and EPBG-LQ were both selected from the same German administrative labor market records. Therefore, both models include several predictors derived from the administrative data, such as age, gender, employment and unemployment history, and education. The response propensity model of EPBG-FQ furthermore contains predictors derived from the sampling frame (e.g., stratum identifiers) and paradata from the CATI interview, such as date and time of a call, interviewer IDs and assessments of the likelihood of a case to participate in the survey which were made by interviewers (Sinibaldi and Eckman 2015). Unfortunately, we cannot include paradata from the online survey in the response propensity model of EPBG-LQ, because paradata is missing for nearly all nonrespondents in this survey.

Using the boosting algorithm and the covariates described above, we predict the response propensity scores, our measure of reluctance, for the respondents and nonrespondents in each dataset. We discuss model performance in the Results section.

5.6.2 Measuring motivated misreporting

We use the differences between the formats (interleafed vs. grouped; go-again vs. how-many), the format effect, as our measure of motivated misreporting. We report results from this analysis in the results section to demonstrate that motivated misreporting is taking place in every survey.

Strictly speaking, the format effect we estimate is not true measurement error, the ϵ term in Figure 5.1. However, comparisons of survey data with administrative records (see Section 5.4) have shown that motivated misreporting, i.e., the format effect, is due to measurement error in the interleafed (go-again) condition. Thus, we can use the format effect to test our hypothesis.

5.6.3 Identification of the relationship between reluctance and motivated misreporting

From the above boosted regression models, we have estimated response propensities for all respondents and nonrespondents. To study the connection between response propensity and motivated misreporting, we split the estimated scores for the respondents into quartiles within each study. The fourth quartiles contain respondents with the highest response propensity scores, i.e., respondents who are the most likely to respond to the survey, given their observed covariates described above. The first quartiles, by contrast, contain respondents with the lowest response propensity scores, i.e., those who responded, but were not likely to do so. For the third part of our analysis, we use only respondents in the fourth and first response propensity quartiles of each dataset. Comparing the format effect between the most likely respondents (the fourth quartile) and the least likely respondents (the first quartile) allows us to study whether reluctant respondents are worse reporters.

One simplistic approach to estimating the connection between response propensity and motivated misreporting is to take the difference in filter questions (looping questions) between reluctant and likely respondents in the interleafed (go-again) format. Yet, this approach will likely produce biased estimates because there may be *true* differences in the behavior measured with the filter questions (looping questions) between reluctant and likely respondents. The two types of respondents differ on many characteristics (recall the covariates of the response propensity models, Table 5.6). For example, reluctant respondents of EPBG-LQ may actually have held more jobs then likely respondents. Without accounting for these true differences in filter and looping questions, estimates of the connection between respondents' reluctance and motivated misreporting may be biased.

To avoid such bias, we extend the analysis of motivated misreporting described in Section 5.6.2 to four subgroups: Reluctant respondents in the interleafed format, likely respondents in the interleafed format, reluctant respondents in the grouped format and likely respondents in the grouped format. We do the same for the looping questions asked in EPBG-LQ (go-again and how-many instead of interleafed and grouped). If response propensity affects motivated misreporting, as predicted by our theory, we should see more misreporting among reluctant respondents than among likely respondents (that is, a larger format effect) in each survey.

That is, we account for true differences in the behavior using the difference in filter (looping) questions between reluctant and likely respondents in the grouped (how-many) format. Because misreporting due to the mechanisms described in Section 5.4 is not possible in the grouped (how-many) format, we assume that any differences between reluctant and likely respondents in the grouped (how-many) format are true differences in behavior. Due to random assignment of respondents to the interleafed and grouped (go-again and how-many) formats, true differences in behavior between reluctant and likely respondents should (in expectation) be about the same in both formats.

In formal terms, we define Y as the number of filter questions triggered, $W \in [0, 1]$ as an indicator of whether a respondent is a reluctant respondent ($W = 1$) or not ($W = 0$) and $I \in [0, 1]$ as an indicator of whether a respondent was interviewed in the interleafed ($I = 1$) or grouped ($I = 0$) format. Under the two assumptions from above (misreporting is only possible in the interleafed/go-again format and the differences between reluctant and likely respondents in the true value, Y^*, are the same in the two formats due to randomization, i.e., $E(Y^*|W = 1, I = 1) - E(Y^*|W = 0, I = 1) = E(Y^*|W = 1, I = 0) - E(Y^*|W = 0, I = 0)$)

and in the absence of a connection between response propensity and misreporting, $E(Y|W = 1, I = 1) - E(Y|W = 0, I = 1) = E(Y|W = 1, I = 0) - E(Y|W = 0, I = 0)$. If there is a connection between respondents' reluctance and misreporting, we should see that the difference between reluctant and likely respondents is larger in the interleafed (go-again) format than in the grouped (how-many) format, due to increased misreporting among reluctant respondents in the former format: $E(Y|W = 1, I = 1) - E(Y|W = 0, I = 1) > E(Y|W = 1, I = 0) - E(Y|W = 0, I = 0)$. In the analysis of EPBG-LQ, Y is the number of events reported and I an indicator whether a person is interviewed in the go-gain ($I = 1$) or how-many ($I = 0$) format.

To better understand the above approach, we can think of it as a difference-in-difference (DiD) model. Difference-in-difference models are commonly used in causal inference settings to derive treatment effects from non-randomized designs (Angrist and Pischke 2009, pp. 221-247). In our case, DiD controls for any true differences in the Y between reluctant and non-reluctant respondents. Just as in the setting described in the previous paragraph, the DiD model is simply the difference of the differences between reluctant and non-reluctant respondents in each format as shown in Equation 5.1.

$$\begin{aligned} DiD = [E(Y|W = 1, I = 1) - E(Y|W = 0, I = 1)] - \\ [E(Y|W = 1, I = 0) - E(Y|W = 0, I = 0)] \end{aligned} \tag{5.1}$$

Estimation of this model is straightforward using a linear regression model, as in Equation 5.2, with intercept β_0, coefficients β_1, β_2, β_3 and residual error term ϵ.

$$Y_i = \beta_0 + \beta_1 I_i + \beta_2 W_i + \beta_3 I_i * W_i + \epsilon_i \tag{5.2}$$

That is, two binary variables (I and W) and their interaction ($I * W$) are included in the model as independent variables. The coefficient of the interaction between these two variables, β_3, is our DiD estimate. Its sign and significance provide the test of our hypothesis that there is a connection between response propensity and motivated misreporting. If β_3 is negative, we interpret this as evidence that there is such a connection and that reluctant respondents show more misreporting. If there is no significant effect, however, we take this as lack of evidence for a connection.

5.7 Results

Presentation of our results proceeds in three steps. First, we present key information on the response propensity models and the estimated response propensity scores for each survey. Second, we analyze whether the data in each survey is affected by motivated misreporting. Third, we present the findings regarding the connection between response propensity and motivated misreporting.

5.7.1 Response propensity models

Table 5.4 shows measures of predictive performance of each response propensity model.[6] Using Youden's J statistic (1950) as a probability cutoff to evaluate model performance, between 71 and 88 percent of respondents are correctly classified as respondents (sensitivity column) and between 64 and 85 percent of nonrespondents are correctly classified (specificity column) (the choice of the probability threshold is only relevant for assessing model performance measures shown in Table 5.4).[7] Taken together, about 65 to 86 percent of all cases are correctly classified ('Accuracy'). Moreover, the area under the receiver operating characteristic curve (AUC) indicates excellent (AUC ≥ 0.8) to outstanding (AUC ≥ 0.9) discrimination in nearly all models, according to the rules of thumb proposed by Hosmer and Lemeshow (2000).[8] R-squared values, that is, the percent of log-likelihood explained by each model, vary between 0.11 and 0.53. Taken together, these performance metrics indicate that the response propensity model built

[6] All models are optimized based on 4-fold cross-validation to guard against overfitting (Hastie, Tibshirani, and Friedman 2009, ch.7).

[7] Many performance measures depend on the probability threshold used to classify cases into a binary classification (e.g., Coussement and Van den Poel 2008). In our study, we choose this threshold as a function of both sensitivity (the true-positive-rate) and specificity (the true-negative-rate). That is, the probability threshold is determined as the cut-point that maximizes both sensitivity and specificity in each survey, as suggested by Youden (1950). Other thresholds commonly applied are 0.5 (Hosmer and Lemeshow 2000) or the observed survey response rate (Tollenaar and Van der Heijden 2013).

[8] The AUC has an appealing interpretation: It is the probability that in a random pair of one respondent and one nonrespondent, the respondent is assigned a higher estimated response propensity than the nonrespondent.

Table 5.4: Performance measures of response propensity models, by survey

Dataset	**Sensitivity**	**Specificity**	**Accuracy**	**AUC**	**McFadden-R^2**
LISS-1	0.82	0.79	0.81	0.89	0.44
LISS-2	0.83	0.79	0.82	0.88	0.39
SOFT	0.88	0.85	0.86	0.94	0.53
EPBG-FQ	0.76	0.75	0.75	0.84	0.28
EPBG-LQ	0.71	0.64	0.65	0.73	0.11

Note: Sensitivity, specificity and accuracy at optimal probability cut-point, as determined by maximal sensitivity and specificity (Youden 1950).

for the SOFT survey discriminates well between respondents and nonrespondents, followed by good predictive performance of LISS-1, LISS-2, and EPBG-FQ. In comparison, the EPBG-LQ response propensity model does not seem to perform as well.

Furthermore, we inspect density plots of the estimated response propensities. Figure 5.2 shows that respondents in each dataset have, on average, higher response propensities than nonrespondents, as expected by the performance metrics discussed above. The covariates included in the response propensity models of LISS-1, LISS-2, SOFT, and EPBG-FQ seem to clearly differentiate between respondents and nonrespondents, a sign that the covariates are good predictors of response to the survey. In line with the findings from Table 5.4, the response propensity model of EPBG-LQ shows less discrimination between respondents and nonrespondents. This may be caused by the limited set of information available to the model.[9]

Table 5.5 shows the ranges of the first and fourth quartile of the estimated response propensity scores in each dataset. Given that we built a unique response propensity model for each dataset, it is not surprising that the range of propensity scores within the first and within the fourth quartiles varies considerably between datasets. Since the value of the response propensity score itself has no meaningful interpretation (Bethlehem, Cobben, and Schouten 2011) and we are only interested in identifying reluctant and likely respondents, differing ranges of response propensity scores across the five studies do not interfere with our analysis.

[9] Unobserved confounders, i.e., predictors of treatment and/or the outcome of interest, are major sources of bias in causal inference settings based on propensity scores where achieving balance on both observed and unobserved confounder is essential to the estimation of unbiased treatment effects. In our study, however, we use the (response) propensity scores only as a tool to identify likely and reluctant respondents and therefore do not need to rely on the strict (and untestable) assumptions necessary for estimation of unbiased treatment effects.

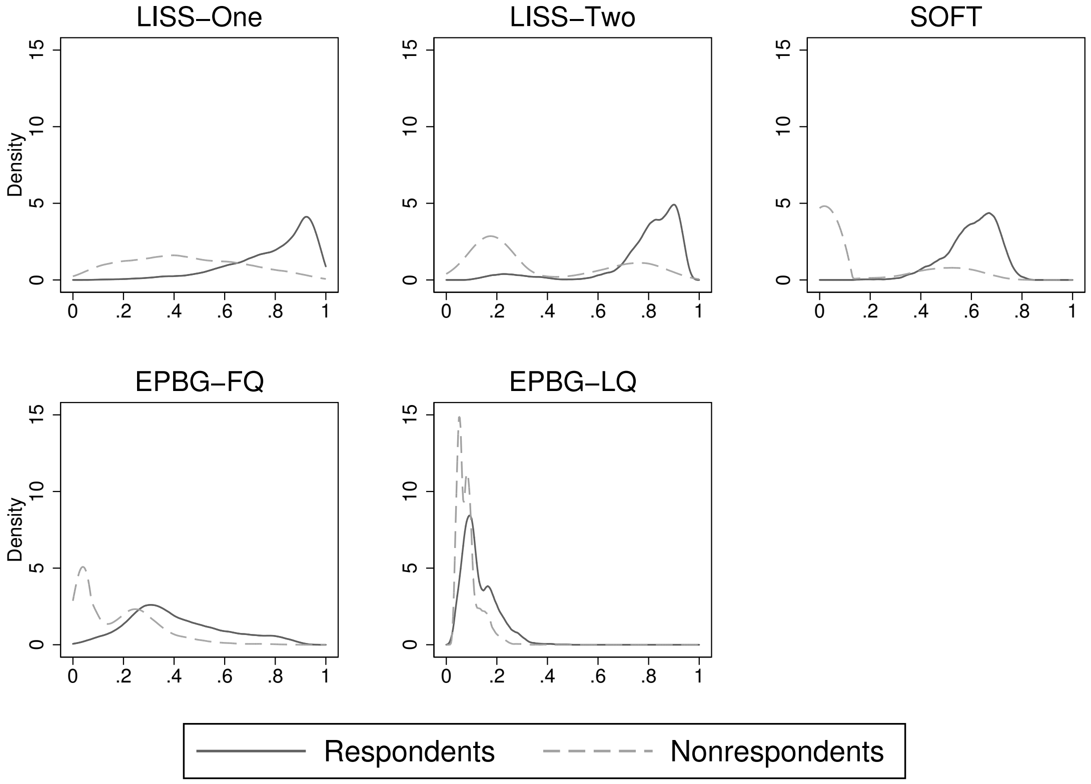

Figure 5.2: Density plots of estimated response propensities, by survey

Table 5.5: Summary statistics of estimated response propensities, by survey

	Quartile						
		1st			4th		
Dataset	Min.	Mean	Max.	Min.	Mean	Max.	n[a]
LISS-1	0.08	0.55	0.70	0.92	0.95	0.99	1,883
LISS-2	0.11	0.55	0.76	0.89	0.92	0.96	1,801
SOFT	0.19	0.47	0.55	0.68	0.71	0.79	152
EPBG-FQ	0.05	0.20	0.28	0.52	0.68	0.89	600
EPBG-LQ	0.03	0.06	0.08	0.17	0.22	0.47	534

[a] *Respondents in first and fourth response propensity quartiles only.*

Table 5.6: Relative influence[a] of most influential predictors of response, by dataset

LISS-1	**LISS-2**	**SOFT**	**EPBG-FQ**	**EPBG-LQ**
Gross household income	Number of filters triggered in wave one	Screening interview started	Appointment made	Age
6.5 %	30.6 %	27.1 %	10.7 %	43.4 %
Year of birth	Gross household income	Date of second contact attempt	Income full-time job	Number of past employers
6.3 %	4.9 %	9.3 %	8.4 %	11.4 %
Age at interview	Net household income	Time of first contact attempt	Year of birth	Type of last spell
5.7 %	4.2 %	5.9 %	8.1 %	8.1 %
Age of household head	Wave one FQ format	Date of first contact attempt	Calls per case	Education
4.7 %	3.9 %	5.8 %	7.5 %	6.3 %
Net household income	Age of household head	Time of second contact attempt	Income part-time job	Stratum
4.5 %	3.0 %	5.6 %	6.7 %	5.3 %
		Number of predictors		
116	120	47	42	15

[a] *Percentage of log likelihood explained by predictor relative to the total log likelihood explained by the model.*

Table 5.6 shows the five most influential predictors of response in each dataset. In LISS-1, sociodemographic information such as the year of birth or having a migration background dominate the response propensity model. In LISS-2, sociodemographic information and covariates collected in LISS-1 have the greatest influence. The response propensity models for SOFT and EPBG-FQ, both telephone surveys, are dominated by paradata collected during call attempts. EPBG-LQ, however, is dominated by sociodemographic information and variables derived from the administrative labor market records. We next turn to the analysis of motivated misreporting in each dataset.

5.7.2 Motivated misreporting

Table 5.7 shows results of the analysis of motivated misreporting in each dataset using *all* respondents. Motivated misreporting is taking place in all four datasets that include filter questions (LISS-1, LISS-2, SOFT, and EPBG-FQ): respondents trigger fewer filters in the interleafed (row one) than in the grouped format (row two). These results support the hypothesis that respondents learn to misreport in the interleafed format. Interestingly, the size of the effect seems to be about one filter question in every dataset (except for SOFT), i.e., misreporting

Table 5.7: Means of filters and events reported, by question format and survey

	Filters triggered				**Events reported**
	LISS-1	LISS-2	SOFT	EPBG-FQ	EPBG-LQ
Interleafed	4.75 (0.05)	4.63 (0.05)	7.39 (0.20)	6.98 (0.10)	-
Grouped	5.58 (0.05)	5.63 (0.06)	7.92 (0.21)	7.81 (0.10)	-
Go-again	-	-	-	-	5.01 (0.10)
How-many	-	-	-	-	6.96 (0.12)
|F-statistic|[a]	125.96	181.26	3.39	32.76	153.34
p-value	0.000	0.000	0.067	0.000	0.000
n	3,767	3,601	304	1,200	1,068

[a] *H_0: Filters triggered interleafed = filters triggered grouped / Events reported go-again = events reported how-many.*
Note: Standard errors in parentheses.

patterns seem to be very consistent across these datasets. The effect in SOFT, significant at the 10% level, however, is only about half a filter question. The smaller effect size could be due to the fact that SOFT is a face-to-face survey, where the physical presence of an interviewer may cause respondents to report more honestly. We observe a similar pattern of motivated misreporting to looping questions in EPBG-LQ.[10] Respondents in the go-again format report almost two fewer events than respondents interviewed in the how-many format, again supporting the hypothesis that respondents learn to misreport. To sum up, we find clear evidence that motivated misreporting is taking place in every dataset: respondents deliberately give false or inaccurate answers to both filter questions and looping questions to avoid follow-up questions and reduce the burden of the survey.

5.7.3 Motivated misreporting among reluctant respondents

In the next step of our analysis, we reduce the analysis sample of each dataset to reluctant (lowest response propensity quartile) and likely respondents (highest response propensity quartile). We then estimate the difference-in-difference models described in Section 5.6.2. Results

[10] The number of loops was limited to a maximum of seven loops for the first section of questions and five loops for the second section of questions (the topic of the questions was randomized, see Section 5.5). Because respondents in the how-many format could report more then seven and five events, we re-code the number of events reported in the this format to match the go-again format, following Eckman and Kreuter (forthcoming).

Table 5.8: Difference-in-difference estimates of the influence of response propensity on motivated misreporting, by survey

	LISS-1	**LISS-2**	**SOFT**	**EPBG-FQ**	**EPBG-LQ**
Interleafed	-0.71***	-0.83***	-0.49	-0.69*	-
(ref. grouped)	(0.14)	(0.14)	(0.61)	(0.29)	-
Go-again	-	-	-	-	-1.05**
(ref. how-many)	-	-	-	-	(0.31)
Reluctant respondent	-0.18	0.06	-0.16	-0.81**	1.46***
	(0.15)	(0.14)	(0.60)	(0.29)	(0.31)
Interleafed*Reluctant	-0.23	-0.12	-0.03	0.22	-
respondent	(0.20)	(0.20)	(0.87)	(0.41)	-
Go-again*Reluctant	-	-	-	-	-1.81**
respondent	-	-	-	-	(0.45)
n[a]	1,883	1,801	152	600	534

*Note: *** p<0.001,** p<0.01, * p<0.05, $^+$ p<0.10. Standard errors in parentheses.*
[a]*Respondents in the first and fourth response propensity quartiles only.*

are reported in Table 5.8. Regarding LISS-1, LISS-2, and EPBG-FQ we find that respondents in the interleafed format trigger fewer filters in the interleafed than in the grouped format (first row), after accounting for respondents' reluctance. Interestingly, the format effect in the SOFT survey seems to disappear, as the coefficient on the interleafed indicator in this model is insignificant. This finding, however, may also be due to the very small sample size of the SOFT survey (recall that we use only half of the respondents in this analysis).

As indicated by the insignificant coefficients on the reluctance indicator, reluctant respondents in LISS-1, LISS-2, and SOFT do not report fewer filters (in the grouped format) than likely respondents. In EPBG-FQ, there seems to be a difference between reluctant and non-reluctant respondents: reluctant respondents report fewer filters than likely respondents ($\hat{\beta} = -0.81, s.e. = 0.29$). Given the assumptions discussed in Section 5.6.3, however, this result is likely due to a true difference in purchasing behavior among reluctant and non-reluctant respondents.

To answer our research question, we check whether these differences in filters triggered are similar in the interleafed format (recall the DiD model described in Section 5.6.3). The absolute size of the interaction effect (third row), our main coefficient of interest, is small and insignificant. That is, the difference in filters triggered between reluctant and likely respondents in the interleafed format is about the same as in the grouped format. Thus, we do not find evidence that motivated misreporting is stronger among reluctant respondents regarding filter questions.

Turning to the connection between response propensity and motivated misreporting to looping questions in the EPBG-LQ dataset, we find that respondents report fewer events in the go-again format ($\hat{\beta} = -1.05, s.e. = 0.31$), as expected. Moreover, there seem to be true differences in the number of events reported in the how-many format between reluctant and non-reluctant respondents, as indicated by the reluctance indicator ($\hat{\beta} = -1.46, s.e. = 0.31$). Contrary to the filter question surveys, however, we find a significant interaction effect ($\hat{\beta} = -1.81, s.e. = 0.45$). That is, the difference between reluctant and likely respondents in the go-again format is much larger than the difference between these two groups in the how-many format. We find evidence for our hypothesis that there is a connection between response propensity and motivated misreporting to looping questions.

Interestingly, reluctant respondents report *more* events than likely respondents in both formats. That is, true differences (given the assumptions discussed in Section 5.6.3) indicate that reluctant respondents had more employers and lived in more places than likely respondents. The difference in the go-again format, however, is larger than the difference in the how-many format, thus, supporting our hypothesis that reluctant respondents are worse reporters (see also Table 5.9 in the appendix).

5.7.4 Robustness

To assess the robustness of the results presented in Table 5.8 to the specification of reluctant and likely respondents, we modify the reluctance indicator. Instead of comparing misreporting between respondents of the first and fourth response propensity quartile, we analyze misreporting of respondents in the first and tenth response propensity *decile*. That is, our definition of reluctance covers only 20% of all respondents (instead of 50%) - the most reluctant 10% and the most likely 10%.

Generally speaking, results of these robustness checks (reported in Table 5.10 and Table 5.10 in the appendix) are in line with the findings presented in the previous section. That is, for LISS-1, SOFT and EPBG-FQ, there is no evidence that motivated misreporting to filter questions is worse among reluctant respondents. Only for LISS-2 do we find a substantial and significant interaction effect ($\hat{\beta} = -0.63, s.e. = 0.31$), that is, misreporting seems to be worse among the most reluctant respondents in LISS-2. Regarding misreporting to looping questions, our substantive conclusions compared to the results reported in the previous section. That is, misreporting is worse among reluctant respondents in EPBG-LQ ($\hat{\beta} = -1.51, s.e. = 0.75$).

To sum up, the results reported in Section 5.8 and the results of the robustness checks do not provide much support to our hypothesis of a connection between response propensity and motivated misreporting to filter questions. However, we find a connection of response propensity and motivated misreporting in the looping questions study, indicating that reluctant respondents are worse reporters to looping questions.

5.8 Discussion

Are reluctant respondents more likely to introduce measurement error, specifically motivated misreporting? Using data from five surveys conducted in different modes and countries, we analyzed the connection between response propensity and two different forms of motivated misreporting, filter questions and looping questions, to answer this question. We estimated response propensities in each dataset using a data mining algorithm that allows us to sidestep the challenge of having to pre-specify a set of predictors of response from all available covariates and their correct functional form. While we did not find much evidence for a connection between response propensity and motivated misreporting to filter questions, we found that reluctant respondents show a stronger tendency to take shortcuts in looping questions than likely respondents.

The finding that reluctant respondents do not misreport more to filter questions than likely respondents is good news for researchers who put extra efforts into achieving high response rates. While high response rates do not necessarily decrease bias due to nonresponse (see the literature reviewed in Section 5.2), we do not find that the extra efforts introduce additional measurement error in terms of increased levels of motivated misreporting.

The nonresponse-measurement error model offers a theoretical explanation of why reluctant respondents may in fact report less accurate data than likely respondents. This model states that the survey reports are a function of the true value and an error term that is determined by the response propensity. Our results, at least for four out of five datasets, do not support this model. We do not believe that this model is wrong, rather, it does not seem to apply to the case of misreporting to filter questions. It may well be that once a sampled person decides to participate in the survey, her motivation or interest in the survey is high enough to give answers as correct as any other respondents. The desire to reduce survey burden may be a good explanation for motivated misreporting, but sampled units with a strong desire to reduce survey burden may simply decide to not participate in the survey at all. This explanation is supported by the fact that the response propensities of even the reluctant respondents lie to the right of most of the response propensities of the nonrespondents in four out of five datasets (see Figure 5.2). In other words, the lowest response propensity cases are in fact nonrespondents, and we are not able to explore the patterns of measurement error among nonrespondents.

Regarding motivated misreporting to looping questions, we found that reluctant respondents are worse reporters because the difference between reluctant and likely respondents in the go-again format is larger than the difference between the two in the how-many format. It may be possible that the burden of answering looping questions is higher than the burden of answering filter questions. In the looping question survey, respondents had to recall all employers they worked for in the last five years and all places they had lived. In most of the filter question surveys, however, respondents were asked for purchases during the last year or month (see Section 5.5 and the text of the questions in the Appendix). Retrieving the information in respondents' minds to answer the latter question seems much easier than remembering all events that had to be reported to the looping question survey. Support for such a difference in burden between the surveys can also be found in the result that the magnitude of misreporting is about one item in the filter question surveys, but almost two items in the looping question survey.[11] In general, it is possible that reluctant respondents tend to give less correct answers only when the burden of answering questions is very high.

While the results from the looping question survey seem to support our theoretical model, there is a major limitation regarding these results. Unlike the response propensity models of the four filter question datasets, which included rich information derived from paradata or respondents' past reporting behavior, the response propensity model of EPBG-LQ could only use covariates derived from the administrative labor market data and information from the sampling frame. Moreover, the most influential predictors of response in the filter question datasets are exactly those that are not available in EPBG-LQ (e.g., whether a case ever refused an interview, the likelihood whether a case would cooperate or the number of contact attempts, see also Table 5.6). In line with these findings, the performance of this model is the worst on all measures. Thus, the response propensity model does not distinguish as well between reluctant and non-reluctant respondents. The findings regarding the connection between response propensity and motivated misreporting in this dataset may be confounded by the limited predictive performance of the response propensity model. Therefore, we consider these results as exploratory findings that require more research and replication with other datasets.

[11] In the latter case, re-coding of the number of events reported in the how-many format to match the maximum number of loops in the go-again format may, however, partially explain the magnitude of the effect (see Section 5.7.2).

5.9 Appendix

5.9.1 Additional results

Table 5.9: Predicted means of filters and events reported, by respondents' reluctance, question format, and survey

		Filters triggered				**Events reported**
		LISS-1	LISS-2	SOFT	EPBG-FQ	EPBG-LQ
Interleafed	Reluctant respondents	4.45 (0.10)	4.64 (0.10)	7.44 (0.47)	6.94 (0.18)	-
	Likely respondents	4.86 (0.09)	4.70 (0.08)	7.63 (0.38)	7.54 (0.21)	-
Grouped	Reluctant respondents	5.39 (0.12)	5.58 (0.12)	7.95 (0.40)	7.42 (0.20)	-
	Likely respondents	5.57 (0.09)	5.52 (0.09)	8.11 (0.48)	8.23 (0.22)	-
Go-again	Reluctant respondents	-	-	-	-	4.90 (0.21)
	Likely respondents	-	-	-	-	5.26 (0.20)
How-many	Reluctant respondents	-	-	-	-	7.76 (0.25)
	Likely respondents	-	-	-	-	6.31 (0.23)
	n[a]	1,883	1,801	152	600	534

[a] *Respondents in the first and fourth response propensity quartiles only.*

Note: Standard errors in parentheses.

Table 5.10: Summary statistics of estimated response propensities, by survey

	Decile						
		1st			10th		
Dataset	Min.	Mean	Max.	Min.	Mean	Max.	n[a]
LISS-1	0.08	0.43	0.56	0.95	0.96	0.99	1,883
LISS-2	0.11	0.32	0.62	0.92	0.93	0.96	1,801
SOFT	0.33	0.40	0.46	0.67	0.72	0.78	152
EPBG-FQ	0.05	0.14	0.20	0.70	0.78	0.89	600
EPBG-LQ	0.03	0.05	0.06	0.22	0.27	0.47	534

[a] *Respondents in first and fourth response propensity quartiles only.*

Table 5.11: Difference-in-difference estimates of the influence of response propensity on motivated misreporting, by survey (lowest and highest decile)

	LISS-1	**LISS-2**	**SOFT**	**EPBG-FQ**	**EPBG-LQ**
Interleafed	-0.67***	-0.79***	-0.15	-0.48	-
(ref. grouped)	(0.23)	(0.22)	(1.07)	(0.47)	-
Go-again	-	-	-	-	-1.32*
(ref. how-many)	-	-	-	-	(0.53)
Reluctant respondent	-0.23	0.61**	0.10	-0.48	1.10*
	(0.23)	(0.23)	(1.13)	(0.48)	(0.51)
Interleafed*Reluctant	-0.34	-0.63*	-0.50	0.06	-
respondent	(0.32)	(0.31)	(1.46)	(0.66)	-
Go-again*Reluctant	-	-	-	-	-1.51*
respondent	-	-	-	-	(0.75)
n[a]	753	721	61	240	213

*Note: *** p<0.001,** p<0.01, * p<0.05, $^+$ p<0.10. Standard errors in parentheses.*
[a]*Respondents in the first and tenth response propensity decile only.*

5.9.2 Filter questions used in LISS-1 and LISS-2

- In the past month, have you purchased coffee for consumption at home?
- In the past month, have you purchased beer or wine for consumption at home?
- In the past month, have you purchased tobacco?
- In the past month, have you purchased children's clothing or shoes?
- In the past month, have you purchased clothing or shoes for yourself?
- In the past month, have you purchased chocolate?
- In the past month, have you purchased medication?
- In the past month, have you purchased flowers?
- In the past month, have you purchased pet supplies?
- In the past month, have you purchased movies on DVD or VHS?
- In the past month, have you purchased music on CD or as MP3s (or other digital formats)?

- In the past month, have you purchased a ticket for a concert, theater performance or a movie?
- In the past month, have you purchased any cleaning supplies for your home?

Follow-up questions

For each yes answer to the above filter questions:

Thinking about your most recent purchase of (fill: item)

- How much did it cost?
  - (Open ended response in Euros)
  - a. Don't know
  - b. Refused
- For whom was it purchased?
  - a. self
  - b. another household member
  - c. someone else
  - d. Don't know
  - e. Refused

5.9.3 Filter questions used in SOFT

Sports section filter questions

- Do you follow professional hockey?
- Do you follow professional basketball?

- Do you follow professional soccer?
- Do you follow professional football?

Sports section follow-up questions

For each yes answer to the above filter questions:

- Do you have a favorite team you support, or do you have no particular favorite team?
  - a. I have a favorite team
  - b. I do not have a favorite team
  - c. cannot say
- When you watch (or attend) (fill: item) games, do you usually do so alone, or with family members or friends?
  - a. alone
  - b. with family members
  - c. with friends
  - d. with both family members and friends
  - e. it depends
- In the last 12 months, how much money have you spent purchasing (fill: item)-team merchandise ?
  - (Open ended response in Dollars)
- When (fill: item) is in season, how many hours per week, on average, do you spend watching (fill: item)?
  - (Open ended response in hours)

- Do you also follow college or high school (fill: item), or only professional (fill: item)?
    - a. college
    - b. high school
    - c. mentioned another league
    - d. only professional

- Were you also a fan as a child or teenager, or did you only become a fan since getting older?
    - a. also fan as child
    - b. fan only since getting older

Clothing section filter questions

- In the last 12 months, have you purchased shoes?
- In the last 12 months, have you purchased jeans or pants?
- In the last 12 months, have you purchased a shirt or sweater?
- In the last 12 months, have you purchased sport clothing?
- In the last 12 months, have you purchased a coat or jacket?
- In the last 12 months, have you purchased a business suit?

Clothing section follow-up questions

For each yes answer to the above filter questions:

- Thinking of the most recent (fill: item) you purchased, was/were it/those for yourself, or for someone else?
    - a. self

  - b. someone else

- How much did the/those (fill: item) cost?
  - (Open ended response in Dollars)

- Was this an impulse purchase or a planned purchase?
  - a. planned
  - b. impulse

- At what kind of store did you buy the/these (fill: item)?
  - a. department or big box store
  - b. local or boutique store
  - c. online
  - d. other

- How comfortable would you feel purchasing such items online?
  - a. very comfortable
  - b. somewhat comfortable
  - c. somewhat uncomfortable
  - d. not at all comfortable

TV section filter questions

- Do you regularly watch soap operas on television?
- Do you regularly watch reality television shows?
- Do you regularly watch late night talk shows?

- Do you regularly watch evening drama shows?
- Do you regularly watch evening scripted comedy shows?
- Do you regularly watch evening news programs?

TV section follow-up questions

For each yes answer to the above filter questions:

- What is the name of your favorite show in this category?
  - (open ended response)
- How many hours per week, on average, do you spend watching such shows?
  - (open ended response in hours)
- Do you believe such programs have educational value?
  - a. yes
  - b. no
- Approximately how many different shows in this category do you watch regularly?
  - (open ended response)

5.9.4 Filter questions used in EPBG-FQ

Clothing section filter questions

- This year, that is in 2011, have you bought a coat or jacket for yourself or for someone else?

- This year, that is in 2011, have you bought a shirt or a blouse for yourself or for someone else?
- This year, that is in 2011, have you bought trousers for yourself or for someone else?
- This year, that is in 2011, have you bought shoes for yourself or for someone else?
- This year, that is in 2011, have you bought sportswear for yourself or for someone else?
- This year, that is in 2011, have you bought swimwear for yourself or for someone else?

Clothing section follow-up questions

- For whom did you purchase this/those (fill: item)? For yourself, a family member, or someone else?
- In what month did you purchase this/those (fill: item)?
- How much did this/those (fill: item) cost?
- How satisfied are you with this/those (fill: item)? Are you very satisfied, somewhat satisfied, somewhat dissatisfied, or very dissatisfied?

Employment section filter questions

- Have you ever held a full-time job? (Note: We explicitly instructed respondents not to include self-employment.)
- Have you ever held a part-time job? (Note: We explicitly instructed respondents not to include self-employment or Mini-Jobs.)
- Have you ever held a so-called Mini-Job, with a payment of 400 Euros a month or less?
- Have you ever received professional training?
- Have you ever received paid practical training?

- Have you ever been self-employed?

Employment section follow-up questions

- From when and until when did you hold your most recent (fill: item)?
- How many hours per week did/do you work in your most recent (fill: item)?
- In what industry was/is your most recent (fill: item)?
- What was your last monthly income at your most recent (fill: item)?

Income section filter questions

- In the year 2010: Did you or another person in your household have income from interest or investment income, e.g., savings, shares, equity funds, or fixed-interest securities?
- In the year 2010: Did you or another person in your household have income from rental property, including leases and subleases?
- In the year 2010: Did you or another person in your household receive a child benefit?
- In the year 2010: Did you or another person in your household receive parental money or a maternity benefit?
- In the year 2010: Did you or another person in your household receive income support?
- In the year 2010: Did you or another person in your household receive unemployment insurance?

Income section follow-up questions

- Which person in your household has received income from (fill: item)? You yourself or another member of your household?
- How often (with what regularity) did your household receive (fill: item)?

- How large was the last amount of income from (fill: item) that your household received in 2010?
- In what month in 2010 did your household first receive income from (fill: item)?

5.9.5 Looping questions used in EPBG-LQ

Employers, how-many format

- For how many employers have you worked in your life so far? Please do not count phases in which you were self-employed.
- Loop through for each reported employer up to 7 (in first section) or 5 (in second section)
  - Please think about your first/second/... employer... When did your employment there begin?
  - When did your employment there end?
  - How many hours did you work there?
  - What was your professional position when you stopped working there?
- If fewer than 7 (5) employers reported: Did you have any other employer you did not think of before?
- If yes:
  - When did your employment there begin?
  - When did your employment there end?
  - How many hours did you work there?
  - What was your professional position when you stopped working there?
- If fewer than 7 (5) employers reported: Did you have any other employer you did not think of before? <continued until respondent said no, or reported about maximum number of events >

Employers, go-again format

- Have you ever been employed? Please do not count phases in which you were self-employed.
- If yes:
  - Please think about your first/second/…employer. When did your employment there begin?
  - When did your employment there end?
  - How many hours did you work there?
  - What was your professional position when you stopped working there?
- Did you have any other employers after that?
  - If yes, loop through follow-up items again
- <continued until respondent said no, or reported about maximum number of events >

Residential locations, how-many format

- In how many places have you ever lived?
- Loop through for each reported location up to 7 (in first section) or 5 (in second section)
  - Please think about your first/second/…place of residence…When did you begin living there?
  - And when did you move away?
  - How many people lived there?
  - Which state is the place in?
- If fewer than 7 (5) locations reported: Have lived anywhere else that you did not yet mention?

- If yes:
  - When did you begin living there?
  - And when did you move away?
  - How many people lived there?
  - Which state is the place in?
- If fewer than 7 (5) locations reported: Have lived anywhere else that you did not yet mention? <continued until respondent said no, or reported about maximum number of events >

Residential locations, go-again format

- Please think about your first/second/…place of residence…
  - When did you begin living there?
  - And when did you move away?
  - How many people lived there?
  - Which state is that in?
- Have you lived anywhere else?
  - If yes, loop through follow-up items again
- <continued until respondent said no, or reported about maximum number of events >

5.9.6 Predictor variables, by survey

Survey	Variables
LISS-1	Gender
	Position in household
	Year of birth
	Age in CBS categories
	Age of household head
	Number of household members

Number of children in household
Household Head lives together with a partner
Civil Status
Domestic situation
Type of household's dwelling
Urban character of place of residence
Primary occupation
Personal gross monthly income (in Euros)
Personal gross monthly income (in Euros), imputed
Personal net monthly income (in Euros) (incl. nettocat)
Personal net monthly income (in Euros)
Personal net monthly income (in Euros), imputed
Personal gross monthly income in categories
Personal net monthly income in categories
Gross household income (in Euros)
Net household income (in Euros)
Highest level of education (without diploma)
Highest level of education (with diploma)
Level of education in CBS categoriess
Household member participates in the panel
Recruitment wave
Ethnic group
Does the household have a simPC?
Count of responses in previous waves
Interviewer ID recruitment interview
Instrument (through which respondent was originally contacted)
Instrument datafile (through which respondent was finally contacted)
Usable address
Contact successful
Minimally posed central question
Complete recruitment interview
Willing to participate in the panel
Registered as panel member
Number of CATI contacts
Number of CAPI contacts
Total number of contact attempts
Interviewer used designated arguments
Life satisfaction (in general)
Statement: Feel good about myself
Confidence in abilities
Follow news on TV or radio
Follow news on the internet
Follow news through free daily paper
Follow news through national or regional newspaper
Follow news hardly or never
Follow news: do not know
Interested in news
Been to cinema in past 12 months
Visited museum in the Netherlands in the past 12 months
Doing voluntary work
Actively taken part in activities of one or more associations or organisations
Doing sports? If yes how many hours a week (on average)
Grade of health at present
Suffer from one or more long-term diseases afflictions or handicaps
Frequency of contact to friends, close acquaintances or family members
Statement: Enough people to fall back on in event of misfortune
Statement: Miss having people around
Statement: Life is meaningless without religion

	Interested in political topics
	Voted in 2006 election
	Which of the two best decribes own view
	Household has computer with internet connection
	Type of internet connection: Cable connection
	Type of internet connection: ADSL
	Type of internet connection: Dial-up connection 1
	Type of internet connection: Dial-up connection 2
	Type of internet connection: Mobile internet
	Type of internet connection: other fast, broadband
	Type of internet connection: do not know 1
	Type of internet connection: do not know 2
	Type of internet connection: Dial-up connection used 1
	Type of internet connection: Dial-up connection used 2
	Type of internet connection: Cable connection
	Type of internet connection: ADSL connection, DSL
	Type of internet connection: Mobile internet
	Type of internet connection: other fast, broadband
	Type of internet connection: do not know
	Household has computer without internet connection
	Computer has: Windows Vista
	Computer has: Windows XP
	Computer has: Windows 2000
	Computer has: Windows 95 or 98
	Computer has: Linux variant
	Computer has: Mac OS (Apple)
	Computer has: other operating system
	Computer has: do not know
	Gender
	Date of birth: day
	Date of birth: month
	Date of birth: year
	Age
	Composition of household
	Household composition
	Age of youngest child in household
	Which of following descriptions best applies to respondent
	Does respondent work
	How many hours a week working in total (in a normal week)
	Highest form of completed education
	How well can respondent make ends meet on household's income
	Estimated gross income
	Born in the Netherlands
	In which country born otherwise
	Farther born in Netherlands
	In which country was farther born otherwise
	Mother born in Netherlands
	In which country was mother born otherwise
	Usually speaking Dutch at home or other language
	Statement: to acquire material posse is one of most important things in life
	Statement: Tolerance better than intolerance
	Statement: Survey research is important for society
LISS-2	Gender
	Position in household
	Year of birth
	Age in CBS categories
	Age of household head
	Number of household members

Number of children in household
Household Head lives together with a partner
Civil Status
Domestic situation
Type of household's dwelling
Urban character of place of residence
Primary occupation
Personal gross monthly income (in Euros)
Personal gross monthly income (in Euros), imputed
Personal net monthly income (in Euros) (incl. nettocat)
Personal net monthly income (in Euros)
Personal net monthly income (in Euros), imputed
Personal gross monthly income in categories
Personal net monthly income in categories
Gross household income (in Euros)
Net household income (in Euros)
Highest level of education (without diploma)
Highest level of education (with diploma)
Level of education in CBS categoriess
Household member participates in the panel
Recruitment wave
Ethnic group
Does the household have a simPC?
Count of responses in previous waves
Interviewer ID recruitment interview
Order of answers
Instrument (through which respondent was originally contacted)
Instrument datafile (through which respondent was finally contacted)
Usable address
Contact successful
Minimally posed central question
Complete recruitment interview
Willing to participate in the panel
Registered as panel member
Number of CATI contacts
Number of CAPI contacts
Total number of contact attempts
Interviewer used designated arguments
Life satisfaction (in general)
Statement: Feel good about myself
Confidence in abilities
Follow news on TV or radio
Follow news on the internet
Follow news through free daily paper
Follow news through national or regional newspaper
Follow news hardly or never
Follow news: do not know
Interested in news
Been to cinema in past 12 months
Visited museum in the Netherlands in the past 12 months
Doing voluntary work
Actively taken part in activities of one or more associations or organisations
Doing sports? If yes how many hours a week (on average)
Grade of health at present
Suffer from one or more long-term diseases, afflictions or handicaps
Frequency of contact to friends, close acquaintances or family members
Statement: Enough people to fall back on in event of misfortune
Statement: Miss having people around

	Statement: Life is meaningless without religion
	Interested in political topics
	Voted in 2006 election
	Which of the two best decribes own view
	Household has computer with internet connection
	Type of internet connection: Cable connection
	Type of internet connection: ADSL
	Type of internet connection: Dial-up connection 1
	Type of internet connection: Dial-up connection 2
	Type of internet connection: Mobile internet
	Type of internet connection: other fast, broadband
	Type of internet connection: do not know 1
	Type of internet connection: do not know 2
	Type of internet connection: Dial-up connection used 1
	Type of internet connection: Dial-up connection used 2
	Type of internet connection: Cable connection
	Type of internet connection: ADSL connection, DSL
	Type of internet connection: Mobile internet
	Type of internet connection: other fast, broadband
	Type of internet connection: do not know
	Household has computer without internet connection
	Computer has: Windows Vista
	Computer has: Windows XP
	Computer has: Windows 2000
	Computer has: Windows 95 or 98
	Computer has: Linux variant
	Computer has: Mac OS (Apple)
	Computer has: other operating system
	Computer has: do not know
	Gender
	Date of birth: day
	Date of birth: month
	Date of birth: year
	Age
	Composition of household
	Household composition
	Age of youngest child in household
	Which of following descriptions best applies to respondent
	Does respondent work
	How many hours a week working in total (in a normal week)
	Highest form of completed education
	How well can respondent make ends meet on household's income
	Estimated gross income
	Born in the Netherlands
	In which country born otherwise
	Farther born in Netherlands
	In which country was farther born otherwise
	Mother born in Netherlands
	In which country was mother born otherwise
	Usually speaking Dutch at home or other language
	Statement: to acquire material posse is one of most important things in life
	Statement: Tolerance better than intolerance
	Statement: Survey research is important for society
	Filter question format in LISS 1
	Number of filters triggered in LISS 1
	Case responded in LISS 1
Soft	Screening interview started
	Call First

	Condition: How-many
	Condition: Grouped
	Condition: Paired
	Primary sampling unit
	Segment
	Transfer case
	Screener completed
	Screener started
	Start disposition code
	Number of contacts total
	Time of 1st contact
	Time of 2nd contact
	Time of 3rd contact
	Time of 4th contact
	Time of 5th contact
	Time of 6th contact
	Time of 7th contact
	Time of 8th contact
	Time of 9th contact
	Time of 10th contact
	Interviewer ID
	Contact mode of 1st contact
	Contact mode of 2nd contact
	Contact mode of 3rd contact
	Contact mode of 4th contact
	Contact mode of 5th contact
	Contact mode of 6th contact
	Contact mode of 7th contact
	Contact mode of 8th contact
	Contact mode of 9th contact
	Contact mode of 10th contact
	Person ever refused interview
	Person refused two times
	Appointment scheduled
	Appointment broken
	Date of 1st contact
	Date of 2nd contact
	Date of 3rd contact
	Date of 4th contact
	Date of 5th contact
	Date of 6th contact
	Date of 7th contact
	Date of 8th contact
	Date of 9th contact
	Date of 10th contact
EPBG-FQ	Appointment scheduled
	Average likelihood to cooperate
	Minimum likelihood to cooperate
	Maximum likelihood to cooperate
	SD of likelihood rating
	Calls per case with likelihood rating
	Share of attempts on weekends
	Share of attempts before 10am
	Share of attempts between 10am and 5pm
	Share of attempts between 5pm and 8pm
	Share of attempts after 8pm
	Stratum
	Always same interviewer

	Share of different interviewers
	Number of calls
	At least one contact without realization
	Gender
	Birthday
	Call accepted
	Call not accepted, busy
	Call not accepted, not connected
	Call not accepted, timeout or no answer
	Call not accepted, port is not reached
	German nationality
	Education
	Case needed phone research after delivery to LINK
	Ever did vocational training
	Vocational training started before 1999
	Ever did a minijob
	Ever did an internship
	Ever worked full time
	Ever worked part time
	Recent employment not full time
	Recent employment not part time
	Received unemployment benefit II in 2010
	Received unemployment benefit I in 2010
	Mean income in last full time spell (in Euros)
	Mean income in part time spell (in Euros)
	Mean income in internship spell (in Euros)
	Mean income in vocational training spell (in Euros)
	Mean income in minijob spell (in Euros)
	Mean income in other spell (in Euros)
EPBG-LQ	Birthday
	Stratum
	Ever did vocational training
	Ever did an internship
	Ever worked full time
	Ever worked part time
	Ever did minijob
	Type of last spell
	Number of moves
	Number of past employers
	Gender
	Highest educational degree
	Vocational degree
	Vocational Training
	German nationality

References

AAPOR (2016). *Standard Definitions: Final Dispositions of Case Codes and Outcome Rates for Surveys.* 9th. Oakbrook Terrace, IL: American Association for Public Opinion Research.

Angrist, J.D. and J.-S. Pischke (2009). *Mostly Harmless Econometrics. An Empiricist's Companion.* Princeton, NY: Princeton University Press.

Bach, Ruben L. and Stephanie Eckman (forthcoming). "Motivated Misreporting in Web Panels". In: *Journal of Survey Statistics and Methodology.*

Bethlehem, Jelke, Fannie Cobben, and Barry Schouten (2011). *Handbook of Nonresponse in Household Surveys.* Hoboken, NJ: John Wiley & Sons, Inc.

Biemer, Paul P. (2001). "Nonresponse Bias and Measurement Bias in a Comparison of Face-to-face and Telephone Interviewing". In: *Journal of Official Statistics* 17.2, pp. 295–320.

Bollinger, Christopher R. and Martin H. David (2001). "Estimation With Response Error and Nonresponse: Food-Stamp Participation in the SIPP". In: *Journal of Business & Economic Statistics* 19.2, pp. 129–141.

Buskirk, Trent D. and Stanislav Kolenikov (2010). "Finding Respondents in the Forest: A Comparison of Logistic Regression and Random Forest Models for Response Propensity Weighting and Stratification". In: *Public Opinion Quarterly* 74.3, pp. 413–432.

Cannell, Charles F. and Floyd J. Fowler (1963). "Comparison of a Self-Enumerative Procedure and a Personal Interview: A Validity Study". In: *Public Opinion Quarterly* 27.2, pp. 250–264.

Coussement, K. and D. Van den Poel (2008). "Churn Prediction in Subscription Services: An Application of Support Vector Machines While Comparing Two Parameter-Selection Techniques". In: *Expert Systems with Applications* 34, pp. 313–327.

Curtin, Richard, Stanley Presser, and Eleanor Singer (2000). "The Effects of Response Rate Changes on the Index of Consumer Sentiment". In: *Public Opinion Quarterly* 64.4, pp. 413–428.

— (2005). "Changes in Telephone Survey Nonresponse over the Past Quarter Century". In: *Public Opinion Quarterly* 69.1, pp. 87–98.

Duan, Naihua, Margarita Alegria, Glorisa Canino, Thomas McGuire, and David Takeuchi (2007). “Survey Conditioning in Self-Reported Mental Health Service Use: Randomized Comparison of Alternative Instrument Formats”. In: *Health Research and Educational Trust* 42.2, pp. 890–907.

Eckman, Stephanie and Frauke Kreuter (forthcoming). “Misreporting to Looping Questions in Surveys: Recall, Motivation and Burden”. In: *Survey Research Methods.*

Eckman, Stephanie, Frauke Kreuter, Antje Kirchner, Annette Jäckle, Roger Tourangeau, and Stanley Presser (2014). “Assessing the Mechanisms of Misreporting to Filter Questions in Surveys”. In: *Public Opinion Quarterly* 78.3, pp. 721–733.

Felderer, Barbara, Frauke Kreuter, and Joachim Winter (2013). “Can We Buy Good Answers? The Influence of Respondent Incentives on Item Nonresponse and Measurement Error in a Web Survey”. In: *Presented at the American Association for Public Opinion Research Annual Conference, Boston, MA.*

Fricker, Scott and Roger Tourangeau (2010). “Examining the Relationship between Nonresponse Propensity and Data Quality in Two National Household Surveys”. In: *Public Opinion Quarterly* 74.5, pp. 934–955.

Groves, Robert M. (2006). “Nonresponse Rates and Nonresponse Bias in Household Surveys”. In: *Public Opinion Quarterly* 70.5, pp. 646–675.

Groves, Robert M., Mick P. Couper, and Sarah Dipko (2004). “The Role of Topic Interest in Survey Participation Decisions”. In: *Public Opinion Quarterly* 68.1, pp. 2–31.

Groves, Robert M. and Emilia Peytcheva (2008). “The Impact of Nonresponse Rates on Nonresponse Bias: A Meta-Analysis”. In: *Public Opinion Quarterly* 72.2, pp. 167–189.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* second. Berlin: Springer.

Hosmer, D. and S. Lemeshow (2000). *Applied Logistic Regression.* New York: Wiley.

Hox, Joop J., Edith de Leeuw, and Hsuan-Tzu Chang (2012). “Nonresponse versus Measurement Error: Are Reluctant Reporters Worth Pursuing?” In: *Bulletin de Méthodologie Sociologique* 113, pp. 5–19.

Keeter, Scott, Andrew Kohut, Carolyn Miller, Robert Groves, and Stanley Presser (2000). "Consequences of Reducing Nonresponse in a Large National Telephone Survey". In: *Public Opinion Quarterly* 64.2, pp. 125–148.

Kessler, Ronald C., Hans-Ulrich Wittchen, Jamie M. Abelson, Katherine McGonagle, Norbert Schwarz, Kenneth S. Kendler, Bärbel Knäuper, and Shanyang Zhao (1998). "Methodological Studies of the Composite International Diagnostic Interview (CIDI) in the US National Comorbidity Survey (NCS)". In: *International Journal of Methods in Psychiatric Research* 7.1, pp. 33–55.

Kreuter, Frauke, Stephanie Eckman, and Roger Tourangeau (forthcoming). "Salience of Burden and Its Effects on Response Behavior to Skip Questions. Experimental Results from Telephone and Web-Surveys". In: *Advances in Questionnaire Design, Development, Evaluation and Testing.* Ed. by P. Beatty, D. Collins, L. Kaye, J. Padilla, G. Willis, and A. Wilmot. Hoboken, NJ: Wiley.

Kreuter, Frauke, Susan McCulloch, Stanley Presser, and Roger Tourangeau (2011). "The Effects of Asking Filter Questions in Interleafed Versus Grouped Format". In: *Sociological Methods & Research* 40.1, pp. 88–104.

Kreuter, Frauke, Gerrit Müller, and Mark Trappmann (2010). "Nonresponse and Measurement Error in Employment Research. Making Use of Administrative Data". In: *Public Opinion Quarterly* 74.5, pp. 880–906.

Lee, Brian K., Justin Lessler, and Elizabeth A. Stuart (2010). "Improving Propensity Score Weighting Using Machine Learning". In: *Statistics in Medicine* 29.3, pp. 237–262.

Martin, Charles L. (1994). "The Impact of Topic Interest on Mail Survey Response Behaviour". In: *Journal of the Market Research Society* 36.4, pp. 327–338.

McCaffrey, Daniel F., Greg Ridgeway, and Andrew R. Morral (2004). "Propensity Score Estimation With Boosted Regression for Evaluating Causal Effects in Observational Studies". In: *Psychological Methods* 9.4, pp. 403–425.

Merkle, Daniel and Murray Edelman (2002). "Nonresponse in Exit Polls: A Comprehensive Analysis". In: *Survey Nonresponse.* Ed. by R. Groves, D. Dillman, J. Eltinge, and R. Little. New York: John Wiley and Sons, pp. 243–258.

Olson, Kristen (2006). “Survey Participation, Nonresponse Bias, Measurement Error Bias, and Total Bias”. In: *Public Opinion Quarterly* 70.5, pp. 737–758.

— (2013). “Do Non-Response Follow-Ups Improve or Reduce Data Quality? A Review of the Existing Literature”. In: *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 176.1, pp. 129–145.

Peytchev, Andy, Emilia Peytcheva, and Robert M. Groves (2010). “Measurement Error, Unit Nonresponse, and Self-Reports of Abortion Experiences”. In: *Public Opinion Quarterly* 74.2, pp. 319–327.

Phipps, P. and D. Toth (2012). “Analyzing Establishment Nonresponse Using an Interpretable Regression Tree Model with Linked Administrative Data”. In: *Annals of Applied Statistics* 6, pp. 772–794.

Rosenbaum, Paul R. and Donald B. Rubin (1983). “The Central Role of the Propensity Score in Observational Studies for Causal Effects”. In: *Biometrika* 70.1, pp. 41–55.

Sakshaug, Joseph and Frauke Kreuter (2014). “The Effect of Benefit Wording on Consent to Link Survey and Administrative Records in a Web Survey”. In: *Public Opinion Quarterly* 78.1, pp. 166–176.

Sakshaug, Joseph, Valerie Tutz, and Frauke Kreuter (2013). “Placement, Wording, and Interviewers: Identifying Correlates of Consent to Link Survey and Administrative Data”. In: *Survey Research Methods* 7.2, pp. 133–144.

Scherpenzeel, Annette (2011). “Data Collection in a Probability-Based Internet Panel: How the LISS Panel Was Built and How It Can Be Used”. In: *BMS: Bulletin of Sociological Methodology / Bulletin de Méthodologie Sociologique* 109, pp. 56–61.

Sinibaldi, Jennifer and Stephanie Eckman (2015). “Using Call-Level Interviewer Observations to Improve Response Propensity Models”. In: *Public Opinion Quarterly* 79.4, pp. 976–993.

Tollenaar, N. and P. G. M. Van der Heijden (2013). “Which Method Predicts Recidivism Best? A Comparison of Statistical, Machine Learning and Data Mining Predictive Models”. In: *Journal of the Royal Statistical Society: Series A (Statistics in Society* 176.2, pp. 565–584.

Tourangeau, Roger, Robert M. Groves, and Cleo D. Redline (2010). “Sensitive Topics and Reluctant Respondents. Demonstrating a Link between Nonresponse Bias and Measurement Error”. In: *Public Opinion Quarterly* 74.3, pp. 413–432.

Tourangeau, Roger, Frauke Kreuter, and Stephanie Eckman (2012). “Motivated Underreporting in Screening Interviews”. In: *Public Opinion Quarterly* 76.3, pp. 453–469.

— (2015). “Motivated Misreporting: Shaping Answers to Reduce Survey Burden”. In: *Survey Measurements. Techniques, Data Quality and Sources of Error.* Ed. by U. Engel. Frankfurt/New York: Campus, pp. 24–41.

Triplett, Timothy, Johnny Blair, Theresa Hamilton, and Yun Chiao Kang (1996). “Initial Cooperators vs. Converted Refusers: Are There Response Behavior Differences?” In: *Proceedings of the American Statistical Association, Survey Research Methods Section*, pp. 1038–1041.

Willimack, D. K., Howard Schuman, B. E. Pennell, and J. M. Lepkowski (1995). “Effects of a Prepaid Nonmonetary Incentive on Response Rates and Response Quality in a Face-to-Face Survey”. In: *Public Opinion Quarterly*, pp. 78–92.

Yan, Ting, Roger Tourangeau, and Zac Arens (2004). “When Less is More: Are Reluctant Respondents Poor Reporters?” In: *Proceedings of the American Statistical Association, Survey Research Methods Section*, pp. 4632–4651.

Youden, W. J. (1950). “Index for Rating Diagnostic Tests”. In: *Cancer* 3, pp. 32–35.

Chapter 6

Conclusion

In the introduction to this dissertation, I pointed out the possibility that repeated measurement in social science (longitudinal) surveys can lead to unintended behavioral consequences. Specifically, answering a set of similar or identical questions can influence respondents' subsequent behavior and/or the way how they report their behavior in subsequent answers. I briefly introduced these forms of change as changes-in-reporting and changes-in-behavior and discussed the circumstances under which they may occur in a cross-sectional survey (survey conditioning) and/or a panel survey (panel conditioning).

Numerous studies have been conducted in the past to uncover the latter type of effects, i.e., changes due to repeated participation in a panel survey. I demonstrated that many of them fall short of addressing three essential methodological challenges: disentangling changes-in-behavior from changes-in-reporting, finding adequate control group data, and adjusting for confounding sources of error. This shortcoming apparently derives from the lack of a unified methodological framework for the study of panel conditioning effects. I filled this gap by first, clearly identifying the three challenges mentioned above, second, discussing the various designs that researchers developed in the past and third, providing a workable methodological framework for the study of panel conditioning effects.

This framework is applied to the analysis of changes-in-behavior, a form of panel conditioning that many previous studies barely accomplished to study with the necessary methodological carefulness. The paper demonstrated that participation in several waves of the German panel study "Labor Market and Social Security" influences respondents' labor market behavior. The true conditioning effect may even be larger than the one estimated as the instrumental variable approach required the inclusion of one- and two-time respondents. Future projects, using similar designs, may overcome this limitation by, for instance, introducing additional (random) variance among first wave respondents through randomly allocated incentives. This variance could then be used as a second instrument to overcome the limitations of this paper (as demonstrated by Fricke et al. 2015).

No evidence was found for a second form of panel conditioning (changes-in-reporting). Ample evidence for survey conditioning in one wave of the LISS panel led to the hypothesis that conditioning would be even stronger in a second wave due to panel conditioning. Although both forms of conditioning share the same theoretical mechanism, only one of the two was present in the data analyzed. As discussed in the paper, the lack of panel conditioning may be due to the short duration of the survey and the low response burden of answering to the questions.

The last empirical paper of this dissertation studied survey conditioning from a new perspective. Due to the widespread, but empirically refuted belief that high response rates lead to low nonresponse bias, many surveys aim for high response rates. This strategy may, however, backfire if reluctant respondents provide data of low quality, i.e., data with increased levels of measurement error. Building on a powerful machine learning algorithm, I identified reluctant respondents via their response propensity. However, there was little evidence that reluctant respondents are worse reporters than those who are the most likely to respond.

All four papers jointly make necessary contributions to the academic research literature on conditioning effects in surveys. At the same time, Chapters 3, 4 and 5 also provide helpful insights for researchers and institutions *collecting* data. Thus, Chapter 4 may help questionnaire designers decide which filter question formats to use and whether using such questions will affect data quality in panel surveys. Furthermore, findings in Chapter 5, suggesting that data quality may not suffer when reluctant respondents are brought into the respondent pool, possibly are good news for data collectors.

Chapter 3 should encourage academic research as well as data collecting institutions to think about ways to avoid panel conditioning in the first place. As mentioned before, there are, to my knowledge, no recommendations regarding the question how to avoid the occurrence of panel conditioning. As soon as panel survey participation affects behavior, however, researchers and data collectors *must* consider the ethical implications of simply asking questions (cf. the discussion in Chapter 2).

Other projects on survey and panel conditioning, in the latter case building on the methodological framework provided in this dissertation, may focus on conditioning effects that occur in data collected with new technologies. For instance, the increasing use of mobile devices in survey research may lead to increases in survey conditioning. People who answer surveys on mobile devices are often less willing to invest as much time in a survey as non-mobile respondents because the former group perceives the act of answering on a mobile device as more burdensome (Mavletova 2013). Thus, they may be more likely to take shortcuts by misreporting.

Furthermore, the findings on misreporting to looping questions among reluctant respondents (Chapter 5) call for replication. Many important predictors of response in the EPBG-LQ survey were not available and the response propensity model may therefore be misspecified, possibly leading to bias in the estimates of the connection between respondents' reluctance and misreporting. Thus, it is necessary for future research to replicate the findings regarding looping questions with other, richer datasets.

References

Fricke, Hans, Markus Frölich, Martin Huber, and Michael Lechner (2015). “Endogeneity and Non-Response Bias in Treatment Evaluation: Nonparametric Identification of Causal Effects by Instruments”. In: *IZA Discussion Paper Series* No. 9428. Available at `http://ftp.iza.org/dp9428.pdf`, last accessed Jan 23, 2018.

Mavletova, Aigul (2013). “Data Quality in PC and Mobile Web Surveys”. In: *Social Science Computer Review* 31.6, pp. 725–743.