mda

methods, data, analyses

Journal for Quantitative Methods and Survey Methodology

Volume 11, 2017 | 1

Data Collection in Panel Surveys

Josef Brüderl & Mark Trappmann (Editors)

Sara Kretschmer & Gerrit Müller — The Wave 6 NEPS Adult Study Incentive Experiment

Ian Plewis et al. — Can Interviewer Observations of the Interview Predict Future Response?

Josef Brüderl et al. — Collecting Event History Data with a Panel Survey

Oliver Lipps & Georg Lutz — Gender of Interviewer Effects in a Multi-topic Centralized CATI Panel Survey

Fabian T. Pfeffer & Jamie Griffin — Determinants of Wealth Fluctuation

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

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

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

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

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

Please register for a subscription via http://www.gesis.org/en/publications/journals/mda/subscribe

Print: Bonifatius Druck GmbH Paderborn, Germany

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

Content

3 Editorial: Data Collection in Panel Surveys
Josef Brüderl & Mark Trappmann

RESEARCH REPORTS

7 The Wave 6 NEPS Adult Study Incentive Experiment
Sara Kretschmer & Gerrit Müller

29 Can Interviewer Observations of the Interview Predict Future Response?
Ian Plewis, Lisa Calderwood & Tarek Mostafa

45 Collecting Event History Data with a Panel Survey: Combining an Electronic Event History Calendar and Dependent Interviewing
Josef Brüderl, Laura Castiglioni, Volker Ludwig, Klaus Pforr & Claudia Schmiedeberg

67 Gender of Interviewer Effects in a Multi-topic Centralized CATI Panel Survey
Oliver Lipps & Georg Lutz

87 Determinants of Wealth Fluctuation: Changes in Hard-To-Measure Economic Variables in a Panel Study
Fabian T. Pfeffer & Jamie Griffin

109 Information for Authors

methods, data, analyses | Vol. 11(1), 2017, pp. 3-6

Data Collection in Panel Surveys

Editorial

Josef Brüderl & Mark Trappmann

During the course of the last decades, panel surveys have gained an increasing importance in the social science infrastructure worldwide and the number of panel studies has risen accordingly, with new panel studies popping up constantly.

The German Data Forum (Rat für Sozial- und Wirtschaftsdaten) has recently identified 77 longitudinal surveys in Germany in the area of social science and economic research, the majority of them panel surveys (Rat für Sozial- und Wirtschaftsdaten forthcoming).

The current success of panel studies is due to at least two specific advantages of this kind of data when compared to cross sectional surveys:

A. The ability to follow individual change across time: The possibility of identifying individual trajectories over the life course is very helpful in many research areas like education, poverty, labour market or public health.

B. The potential for a more rigorous causal argumentation: Unobserved heterogeneity between units of observation is a major threat to causal inference. In panel studies this can be excluded by using within-unit-estimators like fixed-effects estimators which reduce the problem to unobserved heterogeneity within units of observation. In particular, all kinds of treatment evaluation require measurements before the treatment, especially if the treatment is not or cannot be randomized.

However, panel surveys are complex endeavours and in addition to the many error sources known in cross sectional surveys, additional problems arise.

In Germany, the German Data Forum as well as the German National Academy of Sciences (Leopoldina) have just discussed the significance and the challenges of panel surveys and published recommendations (Nationale Akademie der Wissenschaften Leopoldina 2016, Rat für Sozial- und Wirtschaftsdaten forthcoming).

The German Data Forum's recommendations specifically address the requirement of more and more systematic survey methodological research on the growing number of longitudinal surveys while the Leopoldina recommendations emphasize the need for a better methodological qualification of students and early career researchers.

On an international scope, the Panel Survey Methods Workshop series has been initiated and biannual workshops have been held since 2008 with the goal of discussing methodological issues that are specific to panel surveys. Again, on the German national level, a similar workshop series has been started in 2009 and has resulted in 10 meetings with an ever growing number of participants since then.

Thus, panel specific methodological research is currently on a rising trend, but more of this is certainly needed due to the burgeoning number of panel studies. Therefore, this mda special issue on data collection in panel surveys intends to foster this trend by bundling panel methods research papers. The contributions in this issue reflect the broad range of methodological questions that are unique to panel surveys.

Panel attrition – the dropout of former panel members in later waves – is a specific form of nonresponse that can be considered extremely costly. Not only does it threaten to bias results if dropout is non-random. Cases that attrite in wave 2 of a panel can never be used for longitudinal analyses although already considerable costs have been invested in these cases up to this point. Moreover, statistical power decreases continuously as more and more cases from the original sample drop out. Consequently, panel attrition is a major topic in this special issue.

One widely applied instrument to minimize attrition is financial incentives. Different incentives can easily be assigned randomly to respondents. Thus, many studies have been devoted to the effect of incentives. Kretschmer and Müller continue this tradition. They experimentally investigate the effect of switching from promised to prepaid incentives during the course of a panel study. Their outcome is not only the attrition rate, but sample composition and fieldwork effort as well.

A different answer to attrition might be adaptive or responsive fieldwork designs that allow to target respondents at risk of attriting before they attrite and pay them extra attention. Plewis, Calderwood and Mostafa investigate in how far interviewer observations of the interview situation (like whether the respondent enjoyed the interview) might be a useful tool to inform such designs in helping the researcher to detect potential dropouts. Furthermore the potential of these observations in nonresponse correction via imputation or weighting is discussed.

All surveys require a dual inference: From the participants who answer to a certain survey question to the target population of the study (representation) and from the answer to a survey question to a latent or manifest trait of the respondent (measurement).

While the studies on attrition focus on the representation side of panel surveys, the paper by Brüderl, Castiglioni, Ludwig, Pforr and Schmiedeberg focuses on a specific kind of measurement error that is unique to panel surveys: The seam effect that results from inconsistent reporting of events or states at the seam of consecutive waves. The authors demonstrate experimentally how dependent interviewing integrated into an Event History Calendar can be applied to reduce this effect.

Lipps and Lutz in their paper investigate gender of interviewer effects on survey measurement. While this is not a problem specific to panel surveys, panel surveys allow identification of such effects because the same respondent is interviewed repeatedly by different interviewers. This is specifically the case in CATI panel surveys where respondents are distributed quasi randomly across telephone interviewers. Exploiting only within respondent differences the alternative explanation that different interviewers recruit different types of respondents can be ruled out.

The paper by Pfeffer and Griffin is a similar case. They exploit fluctuation in survey reports of net worth of households and investigate to what extent these fluctuations are explained on the one hand by variables measuring socio-economic or demographic changes (hinting at true change in net worth) and to what extent they are explained on the other hand by change of respondents and number of imputed wealth components (hinting at methodological artefacts).

Of course, the papers in this special issue do not address every methodological topic that is relevant to panel surveys. Panel conditioning, the tendency that respondents who have answered repeatedly to certain survey questions show different answer behaviour than first time respondents, is one of the major topics not represented in this special issue. Other interesting topics might have comprised longitudinal weighting, mixing modes in longitudinal surveys, using new media to enhance data collection and panel maintenance and tracking or linking panel surveys to register data or other data sources that enable validation and offer information on attritors. However, we hope that the collection of papers bundled in this special issue makes panel survey research more visible and thereby will spur further research on the methodological foundations of panel surveys.

We thank all the authors and reviewers of this special issue for their commitment and their valuable contributions to this issue.

References

Nationale Akademie der Wissenschaften Leopoldina, acatech – Deutsche Akademie der Technikwissenschaften, & Union der deutschen Akademien der Wissenschaften (2016). The relevance of population-based longitudinal studies for science and social policies. Halle (Saale). https://www.leopoldina.org/uploads/tx_leopublication/2016_Stellungnahme_Laengsschnittstudien_EN.pdf

Rat für Sozial- und Wirtschaftsdaten (forthcoming). Die sozial-, verhaltens- und wirtschaftswissenschaftliche Survey-Landschaft in Deutschland. Empfehlungen des Rat SWD.

The Wave 6 NEPS Adult Study Incentive Experiment

Sara Kretschmer[1,2] *& Gerrit Müller*[2]

1 *Leibniz-Institut für Bildungsverläufe e.V.*

2 *Institut für Arbeitsmarkt- und Berufsforschung (IAB) der Bundesagentur für Arbeit*

Abstract

In wave 6 of the National Educational Panel Study (NEPS) adult starting cohort, an incentive experiment was conducted that randomly switched respondent cash incentives from promised to (partly) prepaid for half of the eligible sample. This research note examines the effects that this change in incentive scheme had on response rates, on sample composition in terms of some key survey variables, and fieldwork efforts by interviewers. We find moderately sized positive effects on overall response rates. The switch in incentive scheme appears to be particularly effective in raising response rates of low educated individuals and those with low reading and mathematics competencies, subgroups that participated underproportionately in prior waves. This differential reaction to the changed incentive scheme therefore leads to a somewhat more balanced sample composition along these dimensions. In line with prior studies, effects on fieldwork efforts such as the number of contact attempts to obtain an interview could be found, but are small in magnitude.

Keywords: NEPS adult study, incentive experiment, nonresponse, bias, sample composition, fieldwork efficiency

1 Introduction

In this research note, we report on the effects of a randomized experiment that switched respondent cash incentives from promised to (partly) prepaid in wave 6 of the National Educational Panel Study (NEPS) adult starting cohort. With regard to interviewer-administered surveys at the household or individual level like the NEPS adult study, it is well known that achieving high response rates is an increasing problem, not only in the German survey environment but also internationally. Several studies document declining response rates over the past decades, both across countries and various survey topics (Atrostic, Bates, Burt & Silberstein, 2001; de Leeuw & de Heer, 2002; Pew Research Center, 2012). As is well known, besides affecting sample size and statistical power of a study, the issue is that unit nonresponse may lead to nonresponse bias when sample members' characteristics differ between respondents and nonrespondents (Schnell, 1997; Groves et al., 2006; Groves & Peytcheva, 2008; Bethlehem, Cobben & Schouten, 2011). That is, depending on the nature of the relation between sample members' individual likelihood to respond and key survey variables, unit nonresponse may induce selection bias into substantive analyses based on data of the realized sample only. Considering the initial waves of the NEPS adult study, there appears to be evidence of selective initial nonresponse and attrition related to educational attainment and basic competencies. In particular, lower educated individuals are less willing to respond both in the first wave and in consecutive panel waves (Zinn, Aßmann & Würbach, 2015). In a similar vein, Kleinert, Christoph & Ruland (2015) report that participants with lower mathematics and reading proficiency attrite from the panel more frequently.

In an effort to keep unit nonresponse and subsequent attrition low, the NEPS adult study offered (conditional) cash incentives right from its inception. The use of cash incentives for respondents has become common practice in most academic surveys in Germany in recent years (e.g. Blohm & Koch 2013; Börsch-Supan, Krieger & Schröder 2013; Blom, Gathmann & Krieger 2015). Pforr et al. (2015) currently offer the most comprehensive overview of incentive effects on response rates and nonresponse bias for Germany, based on eight major cross-sectional and panel surveys (ALLBUS, GIP, NEPS, PAIRFAM, PASS, PIAAC, SHARE and SOEP; Ibid. p.2, for more details on the cited surveys.). However, at the time of their writing, Pforr et al. (2015) only considered evidence from a comparatively small *pilot study* (infas, 2009) to the actual NEPS adult study. In that regard, this research note seeks to complement previous findings and is the first to report on the effects of monetary

Direct correspondence to
Sara Kretschmer
Leibniz-Institut für Bildungsverläufe e.V., Wilhelmsplatz 3, 96047 Bamberg
E-mail: sara.kretschmer@lifbi.de

respondent incentives for the *main study* of the NEPS adult cohort. Thereby, it is also the first to document the wave 6 incentive experiment: In waves 1-5, respondent cash incentives were always provided conditionally on the interview. As we shall explain in more detail below, in wave 6, an experiment was conducted that randomly switched respondent cash incentives from promised to (partly) prepaid for half of the eligible sample.

Against this backdrop, the aim of this research note is threefold: First, we are going to examine how the partial switch to prepaid incentives affected wave 6 response rates, overall, and differentiated by prior wave response status. Second, given the initial nonresponse and attrition biases in terms of educational attainment and competencies referred to above, we explore how this intervention affected sample composition along these particularly relevant (for NEPS) dimensions.[1] Third, given the ever increasing costs associated with fieldwork, especially in face-to-face mode, we investigate how the changed incentive scheme affected fieldwork efficiency as measured by the number of contact attempts per interview and speed of survey response.

As this note is deliberately exploratory in nature, we shall only briefly draw on some common theoretical perspectives related to "social exchange" (Dillman, Smyth & Christian 2014) and "leverage-salience" (Groves, Singer & Corning 2000) to identify potential mechanisms driving the (changed) participation behavior in response to the changed incentive scheme. Since the NEPS adult cohort study has used conditional cash incentives from the beginning, the key change to consider theoretically is the move towards prepaying: (part of) a promised payment for participation is being turned into a payment, or token of appreciation, provided in advance. Viewing the request for survey participation as a specific form of social interaction and exchange, the move towards unconditional giving may evoke behavioral "norms of reciprocity" (Gouldner, 1960). That is, recipients of the prepaid incentive may feel obligated to "return the favor" and respond positively to the subsequent survey request. Especially for individuals on the brink of (non) participation this mechanism may override other -negatively valued- aspects of the survey request, "tilting the scale" in favor of participation (c.f. Groves et al., 2004; p. 177). For example, one may think of those generally uninterested in the survey topic (here in the NEPS context, probably the lower educated), or one may think of those sensing a particularly high burden or time demands of participation (as potentially manifested in temporarily dropping out in a wave before). However, whether the described reciprocity mechanism is indeed that powerful, and how exactly it would affect various subgroups differentially, is difficult to settle a priori.

[1] The authors of this research note were not involved in the design of the experiment. Given that the intervention was not targeted at particular subgroups but applied equally to the full eligible sample, we assume that the primary goal was to increase survey participation by and large.

Recipients of prepaid incentives may just as well not conform to norms of reciprocity, or even feel pressured into the survey, questioning the legitimacy of the survey sponsor altogether (e.g. Börsch-Supan et al., 2013). For individuals who attach a high importance to these aspects of a survey request, prepaying may actually push against participation.

The remainder of this note is structured as follows. In the next section we shall briefly refer to the empirical literature on the effects of respondent incentives in cross-sectional and panel surveys. After this, we introduce a few relevant survey design features of the NEPS adult study, describe the wave 6 incentive experiment, and define our analysis sample. In what follows, we present the effects of prepaid incentives on overall wave 6 survey participation (differentiated by respondents' wave 5 outcome) and then turn to our key empirical findings concerning sample composition in terms of educational background and competence test results. Finally, we investigate the effects of prepaid incentives on fieldwork efficiency and conclude with a brief summary of our findings.

2 Some Previous Research on Incentives and Survey Participation

There is a considerable empirical literature on the effects of respondent incentives on participation, based on cross-sectional and panel surveys of varying topics, conducted in different modes, by various survey sponsors and fieldwork agencies, across several countries. Given the brevity of this research note, we abstain from an extensive literature overview here. In that regard, Singer, van Hoewyk, Gebler, Raghunathan and McGonagle (1999) and Laurie and Lynn (2009) both provide comprehensive overviews of the international literature, the former focusing on respondent incentives in cross-sectional surveys, the latter on longitudinal surveys. As mentioned above, Pforr et al. (2015) recently summarized the evidence for Germany, concluding that most of the international findings carry over to the German survey environment.

In a nutshell, past empirical research on the effects of respondent incentives in interviewer-administered surveys typically finds that incentives increase response rates, that monetary incentives are more effective than non-monetary incentives, and that prepaid incentives affect response rates more strongly than conditional incentives (e.g. Singer et al., 1999; Singer, 2002; Yu & Cooper, 1983; Willimack, Schumann, Pennel & Lepkowski 1995; Ryu, Couper & Marans, 2005). In addition, there are studies suggesting that large incentives increase response rates more than small incentives, albeit at a decreasing rate (e.g. Mercer, Caporaso, Cantor & Townsend, 2015; Scherpenzeel & Toepoel, 2012; Rodgers, 2011).

When incentives are introduced at later waves of panel surveys it is usually found to generate much smaller increases in response rates than similar incentives would yield in cross-sectional surveys, or initial waves of panel surveys (e.g. Laurie & Lynn, 2009; Laurie, 2007; Jäckle & Lynn, 2008). One likely reason is that the panel attrition, which is typically largest in the early waves, has left a fairly cooperative sample that responds rather little to later changes in the incentive scheme. However, one subgroup that typically does react quite strongly to introducing (or increasing) incentives in panel surveys are nonrespondents at the previous wave (e.g. Zagorsky & Rhoton, 2008; Rodgers, 2011).

Relatedly, there are studies suggesting that incentives may be effective in boosting participation of certain demographic groups ordinarily underrepresented, such as people with lower income, ethnic minority status (e.g. James, 1997; Mack, Huggins, Keathley & Sundukchi, 1998) or with low education status (e.g. Berlin et al., 1992; Ryu et al., 2005). However, overall, the evidence is somewhat more mixed than the selected references suggest. For instance, in their meta-analysis Singer et al. (1999) also refer to a number of studies showing no favorable effect of respondent incentives on sample composition at all (Ibid. p. 224-225).

Finally, incentives may affect fieldwork efficiency by reducing the number of calls an interviewer has to make in order to obtain an interview. For example, James (1997) and Rodgers (2002) both find that providing cash incentives may lead to a reduction in the number of calls per completed interview, although the orders of magnitude are rather small. Similarly, in a recent study based on the German General Social Survey (ALLBUS) Blohm and Koch (2013) found a slight reduction in the average number of contact attempts per completed interview by the use of monetary incentives. Mann, Lynn and Peterson (2008) point out that incentives may positively affect early survey response and response speed, thereby increasing fieldwork efficiency through the reduction of intense (and costly) follow-up efforts that would otherwise be necessary.

3 Design and Sample of the NEPS Adult Study

The NEPS is the largest longitudinal study for educational research in Germany. It was established in 2009 for the purpose of collecting survey data about learning environments, educational decisions and returns to education over the entire life-course (Blossfeld & von Maurice, 2011). Furthermore, one of the core issues is to assess the development of competencies, such as reading, basic mathematics or ICT proficiency, and their repeated measurement (Allmendinger et al., 2011). In order to provide data across several periods of life as soon as possible, the NEPS fielded six separate starting cohorts of different age groups. The NEPS adult study, on which we report here, comprises the oldest age groups born between 1944 and

1986, with a questionnaire focused on adult education and the development of competencies in adulthood. The NEPS adult study is conducted annually since 2009. All sample members were drawn from resident registers (Einwohnermelderegister) run by the municipal residents' registration offices, and represent individuals living in private households in Germany born between 1944 and 1986 (Zinn et al., 2015). The first wave of the NEPS adult cohort comprises participants of the 2007/08 prequel study "Working and Learning in a Changing World" (ALWA) born between 1956 and 1986.[2] All respondents to the ALWA study who agreed to be contacted for further interviewing were included in the gross sample of the NEPS adult cohort initial wave in 2009/10. This core wave 1 sample was again supplemented by two additional samples: first-time participants in the same age range as the original ALWA sample (boost sample) and older respondents born between 1944 and 1955 (augmentation sample). In NEPS wave 3, another refreshment sample was added consisting of all birth cohorts from 1944 to 1986 (for further details, be referred to the documentation by the Leibniz-Institut für Bildungsverläufe e.V., 2015).

In the initial wave, respondents are asked about their social and migration background as well as their educational, job and family history retrospectively. These retrospective data are continuously updated in subsequent waves. Moreover, respondents answer questions about their social and cultural capital, health, well-being and social and political participation (Allmendinger et al., 2011). All data are collected in a mixed-mode design with computer assisted telephone interviews (CATI) and computer assisted face-to-face interviews (CAPI). In the initial wave and in every *odd* wave, computer assisted telephone interviewing is the default mode. In *even* waves respondents are asked to additionally take part in competence assessment with paper and pencil, or computer-based. In these waves, face-to-face interviewing is the default mode. In each wave, a small number of interviews is conducted in Turkish or Russian, mainly in telephone mode. If respondents are hard to contact or initially refuse participation in either mode, the study design allows for a mode switch. Participants who do not respond in one or more waves remain in the sample and keep being contacted in subsequent waves. Only those who eventually cannot be located and contacted anymore, or those who explicitly refuse to further participate ("hard refusals") are excluded from the sample.

Up to and including wave 6, three "rounds" of competence assessment have been administered to participants in the even waves. In order to keep the overall burden low, wave 2 sample members were randomly assigned to one of three groups: reading assessment only, mathematics assessment only, both assessments. In wave 4, all sample members who had entered the study in the first NEPS wave were asked to take part in science literacy and information and communication technology (ICT) assessment. Respondents who had entered in the third wave

[2] For details on the ALWA survey, which has been conducted by the Institute for Employment Research (IAB), be referred to Antoni et al. (2010).

(refreshment sample) were asked to take part in reading assessment. In wave 6, the competence assessment includes measurements on listening comprehension at word level and general cognitive functions for all sample members.

3.1 The Incentive Experiment in Wave 6

Based on evidence from the NEPS pilot study (infas, 2009; Pforr et al., 2015), respondents of the main study were offered conditional cash incentives right from the beginning. In wave 1, the NEPS adult study started out with a 10€ cash incentive, which was temporarily raised to 50€ in the second half of the wave 1 fieldwork period due to low initial response. In wave 2, the incentive was increased to 25€ cash conditional on the interview throughout. From wave 3 to wave 5, the incentive was again lowered somewhat to 20€ cash conditional on the interview. In wave 6, the mentioned randomized split-half experiment was used to test the effects of switching to prepaid incentives: one group kept receiving 20€ conditional on the interview as in previous waves (control group). The other group received 10€ unconditionally with the advance letter and another 10€ conditional on the interview (treatment group). The randomization happened at the respondent level. That is, in principle, each one of the 255 CAPI interviewers initially working the sample had cases with and without prepaid incentives. The experiment was run "half blind", that is interviewers knew the incentive status of individual sample members, but each potential respondent was uninformed about the experiment.[3]

3.2 Analysis Sample & Data

Our analysis is based on all sample members eligible for a wave 6 interview. We exclude foreign language interviews because these cases were not part of the randomized experiment. This leads to an analysis sample of 12,280 cases. As just explained, about half of them received postpaid incentives only (n= 6,146) as in previous waves, while the other half received 10€ with the advance letter plus another 10€ conditional on the interview (n= 6,134). To measure the effects of pre-

[3] Given the half-blind design, one may wonder whether interviewers worked cases with prepaid incentives first, thereby implicitly driving some of the differences in outcomes. Similar to Börsch-Supan at al. (2013) for SHARE, working with the same survey agency and prepaid incentives, we did not find evidence for that. The average number of days until the first contact attempt (after a case is being released to an interviewer) is equal across the two incentive conditions.

paid incentives, we employ survey data from wave 1 to wave 5.[4] In addition we also use wave 6 call record data provided by the fieldwork institute to identify the final outcome[5] and analyze fieldwork efficiency.

4 Results

First, we shall briefly present our findings on the effect of prepaid incentives on overall response, contact, and refusal rates. We then differentiate further and evaluate the effects separately for wave 5 respondents and nonrespondents, distinguishing among several reasons for previous wave nonresponse. In what follows, we focus on whether the changed incentive scheme differentially affected the participation of various subgroups in terms of education status and competencies. We find this a good starting point for identifying relevant (for NEPS) selection effects, rather than looking at some arbitrary set of sociodemographic variables that may in the end only be weakly related to the substantive variables of interest.[6] Finally, we investigate the effects of prepaid incentives on some indicators of fieldwork efficiency.

4.1 The Overall Effect of Prepaid Incentives in Wave 6

Overall response rates have been fairly constant, levelling off between 77% and 79% (RR1 following the standard definitions of The American Association for Public Opinion Research (2015) in the waves prior to the experiment. Concerning wave 6, we find that for sample members with (partly) prepaid incentives response rates are somewhat higher (80%) as compared to those with postpaid incentives only (78%). The difference of about 2 percentage points is not very large, yet statistically significant[7] (p-value 0.006). About 15% amongst sample members with postpaid incentives refuse participation, while only 13% with prepaid incentives refuse. This reduction in refusals essentially accounts for the overall 2 percentage point differ-

[4] This paper uses data from the National Educational Panel Study (NEPS): Starting Cohort Adults, doi:10.5157/NEPS:SC6:5.1.0. From 2008 to 2013, NEPS data was collected as part of the Framework Program for the Promotion of Empirical Educational Research funded by the German Federal Ministry of Education and Research (BMBF). As of 2014, NEPS is carried out by the Leibniz Institute for Educational Trajectories (LIfBi) at the University of Bamberg in cooperation with a nationwide network.

[5] The final outcomes recorded in the call record data may marginally differ from outcomes reported in the final method report published by the survey institute.

[6] However, for completeness and comparison with other studies, we have included a table in the appendix showing response rates (treatment vs. control group) for a whole set of variables typically considered (Appendix, Table A1).

[7] Proportions compared with two sample t-tests taking into account clustering at the PSU level (municipalities).

Table 1 Wave 6 Response Rates by Wave 5 Outcome (N= 12,280)

	postpaid			(partly) prepaid			col	
	(n= 6,146)	Interview		(n= 6,134)	Interview		(2) - (1)	p-value
W5 respondents	*5,295*	85.1%	*4,508*	*5,217*	87.4%	*4,561*	2.3%	0.001
W5 nonrespondents	*851*	34.1%	*290*	*917*	38.6%	*354*	4.5%	0.033
Refusals		31.3%	*137*		34.1%	*151*	2.8%	0.363
Noncontacts		44.8%	*26*		51.9%	*41*	7.1%	0.409
Appointments		28.2%	*51*		40.5%	*85*	12.3%	0.006
Other nonrespondents		43.7%	*76*		41.6%	*77*	-2.1%	0.678

ence. Concerning overall contact rates of the wave 6 gross sample, we do not find any effect of prepaid incentives.

4.2 The Effect of Prepaid Incentives by Wave 5 Outcomes

Looking at the effects on response rates in more detail, we find that both response propensities and incentive effects on participation are very different depending on the previous wave outcome. For those who did not participate in the previous wave, we observe an average wave 6 response rate of about 36% as compared to 86% for wave 5 respondents. Looking at the differences between treatment and control cases within these two groups, we see an increase of 2.3 percentage points for wave 5 respondents, and of about twice that size (4.5 percentage points) for those not responding in wave 5 (Table 1).

Differentiating by the reasons for nonresponse within the group of wave 5 drop-outs, we see that the changed incentive scheme is not particularly effective in bringing back prior “refusers” into the sample. The 2.8 percentage point increase is statistically insignificant and also somewhat below the group average of 4.5 percentage points. Rather, those with an appointment as final status in the prior wave react overproportionately strong to the change in incentives. The increase in response rate of 12.3 percentage points is comparatively large and statistically significant. There is also some indication that those who could not be successfully contacted in the prior wave react positively to the prepaid incentive (an increase of 7.1 percentage points). However, we have to interpret these findings with some caution, as the number of cases in these categories is rather small.

Table 2 Wave 6 Response Rates by Educational Background (N= 12,266)

	postpaid			(partly) prepaid			col	p-
	(n= 6,137)	Interview		(n= 6,129)	Interview		(2) - (1)	value
Lower/middle secondary schooling	*310*	64.8%	*201*	*301*	72.8%	*219*	7.9%	0.049
Lower/middle secondary schooling + vocational training	*3,007*	77.7%	*2,337*	*2,984*	79.7%	*2,378*	2.0%	0.060
University-entrance diploma	*1,036*	77.3%	*801*	*1,061*	78.4%	*832*	1.1%	0.541
University/ of applied science	*1,724*	81.6%	*1,406*	*1,727*	84.1%	*1,453*	2.6%	0.034
No degree	*60*	78.3%	*47*	*56*	57.1%	*32*	-21.2%	0.020

4.3 The Effect of Prepaid Incentives on Lower Educated Sample Members

We also examined the effects of prepaid incentives on one of the major NEPS focus variables, the educational attainment of participants. The response rate of individuals with lower or middle secondary schooling degree, and without a vocational training certificate, is about 8 percentage points higher in the experimental treatment condition (Table 2). The increases in the remaining categories are between 1.1 and 2.6 percentage points and therefore close to the overall effect of prepaid incentives of about 2 percentage points[8].

The overproportionate increase in response of the low educated counteracts, at least somewhat, existing biases. Put differently, “representativity” (in the sense of Bethlehem et al., p. 181) with respect to educational attainment is increased, as the response propensities over the four educational degree categories are more equal in

[8] There is a small number of sample members without any schooling or vocational degree (“no degree”). For this group we make the somewhat odd finding of a 21 percentage point decrease in response rates with prepaid incentives. Individuals in this group are on average somewhat older as compared to the rest of the sample and with a migration background more often. In light of the small number of observations we find it difficult to further interpret this finding.

the prepaid than in the postpaid incentive condition.[9] Although the magnitude of this balancing effect is not overly large, it contributes to an enhanced sample composition along the dimension of educational attainment.

4.4 The Effects of Prepaid Incentives on Sample Members with Lower Reading Test Scores

Another core issue of the NEPS adult study is the measurement of participants' competencies, in particular those related to educational success and labor market outcomes like reading or mathematics proficiency (Allmendinger et al., 2011). For our empirical analysis of the NEPS wave 6 incentive experiment we focus on test scores for reading proficiency. This is because reading tests have been administered to the majority of respondents in previous waves, whereas mathematics tests have so far been carried out only for two subsamples of the NEPS adult cohort.[10] In wave 8, reading assessment will be repeated for the first time.

For our analysis of the incentive experiment we distinguish between sample members with no, lower, middle and higher reading test results. For this purpose, we use the available reading competence scores (Pohl & Carstensen, 2012) from prior waves for all cases that participated in the assessment and sort them into three categories, each containing a third of the sample. Those who participated in the respective prior wave but who refused or aborted the competence assessment (or who have been switched to telephone mode) are classified as "no test". Looking at Table 3, we observe that the latter group reacts particularly strong to the changed incentive scheme (5.7 percentage point increase).

One mechanism could be that these respondents sensed an especially high burden of competence assessment participation in previous waves, which are -in part- compensated for by the prepaid incentive when it comes to participation in the current wave. Similarly to the results for educational attainment, we also find here that sample members with the lowest test scores show the largest increase in response rates in reaction to the changed incentive (3.3 percentage points). The effect is on the brink of significance at the 5% level and again not very large. Still, the direction is towards a more balanced sample in terms of reading competence

[9] Note, that the concept of "representative" response is always defined with respect to a selected (set of) variable(s). In practice, one calculates the variance of (estimated) individual response probabilities across the various categories of the chosen variable(s). Intuitively: if there turns out to be little variation in the estimated probabilities across categories, this is taken as evidence against a strong relation between (non)response and the characteristic under consideration. Note, too, that our example of considering variation of average response propensities across educational attainment categories is closely related to what Bethlehem et al. (2011) call an unconditional partial R-indicator.

[10] There are only 5,645 cases with mathematics scores, which is less than half the number of cases in our analysis sample.

Table 3 Wave 6 Response Rates by Reading Proficiency (N= 9,295)

	postpaid			(partly) prepaid			col	
	(n= 4,650)	Interview		(n= 4,645)	Interview		(2) - (1)	p-value
No test	*891*	68.1%	*607*	*841*	73.8%	*621*	5.7%	0.007
Lower tercile	*1,252*	77.2%	*966*	*1,264*	80.5%	*1,017*	3.3%	0.061
Middle tercile	*1,272*	83.7%	*1,065*	*1,254*	84.6%	*1,061*	0.9%	0.542
Higher tercile	*1,235*	87.0%	*1,074*	*1,286*	86.5%	*1,112*	-0.5%	0.709

scores, thereby again counteracting somewhat the existing biases along this dimension. For the restricted sample with mathematics test scores, we found qualitatively similar -yet even weaker- results as compared to reading test scores (Appendix, Table A2).

4.5 The Effects of Prepaid Incentives on Fieldwork Efficiency

In this section we explore the effects of the switch in incentive scheme on the number of contact attempts per interview as well as on the speed of survey response measured in days since the beginning of the fieldwork. Since nonresponse in the previous wave indicates that sample members may be hard to contact and/or less willing to cooperate, we analyze the effects separately by wave 5 response status.

In our call record files for wave 6, we observe a total of 30,369 contact attempts with sample members being assigned to postpaid incentives, and 30,137 contact attempts with cases being assigned to prepaid incentives. The overall workload, as measured by the total number of attempts, hence does not differ much. However, comparing the average number of contact attempts necessary to obtain an interview, we find that prepaid incentives may in fact reduce the number of unproductive contact attempts.[11] This holds at least for sample members that did not respond in wave 5 (see Table 4). For this group, we find a reduction from, on average, 4.5 contact attempts to 3.9 contact attempts. In relative terms, this amounts to a reduction of almost 13% after all. Amongst sample members that did respond in wave 5 there was no significant difference.

[11] From a cost perspective, note, that 90% of all contact attempts per completed interview were personal contact attempts by F2F interviewers since the default mode in wave 6 was CAPI. Out of the 9.713 wave 6 interviews only about 6% (582) were conducted by telephone.

Table 4 Number of Contact Attempts before Interview by Wave 5 Outcomes (N= 9,712)

	postpaid			(partly) prepaid				
	n	Contact attempts	Contact attempts (average)	n	Contact attempts	Contact attempts (average)	col (2) - (1)	p-value
All	*4,797*	14,560	3.04	*4,915*	14,867	3.02	-0.01	0.856
W5 respondents	*4,507*	13,256	2.94	*4,561*	13,478	2.96	0.01	0.821
W5 nonrespondents	*290*	1,304	4.50	*354*	1,389	3.92	-0.57	0.074

Table 5 Number of Days before Interview by Wave 5 Outcomes (N= 9,712), Median

	postpaid		(partly) prepaid			
	n	median	n	median	col (2) - (1)	
W5 respondents	*4,507*	109	*4,561*	105	-4***	(1.466)
W5 nonrespondents	*290*	135	*354*	125	-10**	(4.638)

Standard errors in parentheses; based on median regression analysis
*** p<0.01, ** p<0.05, * p<0.1

For speed of survey response we look at the average (median) number of days between the beginning of the fieldwork and the realized interview. We find a reduction from 109 to 105 days until the interview for wave 5 respondents (see Table 5).

For sample members that did not respond in wave 5, prepaid incentives reduce the number of days until the interview even more, from 135 to 125 days. That shows that sample members respond somewhat faster when receiving prepaid incentives.

5 Conclusion

Summing up, the experimental switch of respondent cash incentives from promised to (partly) prepaid in the wave 6 NEPS adult study certainly brought about positive effects on response rates, sample composition in terms of some key survey variables, and fieldwork efforts. All our findings are in line with the existing literature on incentive effects briefly discussed in the beginning. Nevertheless, the magnitudes were always of rather modest size. Given that the change to the existing incentive scheme can also be considered fairly moderate, this aligns well. In the end, the shift from postpaid to prepaid respondent incentives was implemented only halfway, as only 10€ of the 20€ available per case were now offered unconditionally. In light of our findings for this "partial" move towards prepaid incentives, one might consider switching to prepaid incentives (for panel cases) entirely in future; although no clear predictions about the various effects of such a move are borne out by our analyses. That said, we agree with the conclusion of Blohm and Koch (2013) that changing respondent incentives is -after all- only *one* way of altering survey operations. Deciding what is the most (cost) effective way of raising response rates and affecting sample composition favorably would, among others, necessitate detailed insights into the true cost structure of fieldwork agencies in combination with further experiments. Despite the limitations in terms of generalizability often associated with such single experiments, we believe that findings for large scale surveys should be documented and made available to other researchers and survey practitioners. In that respect, this note adds one piece of evidence to the literature, especially for the German case as recently summarized by Pforr et al. (2015).

References

Allmendinger, J., Kleinert, C., Antoni, M., Christoph, B., Drasch, K., Janik, F., Leuze, K., Matthes, B., Pollak, R. & Ruland, M. (2011). Adult education and lifelong learning. In H.-P. Blossfeld, H.-G. Roßbach & J. von Maurice (Eds.), *Education as a lifelong process. The German National Educational Panel Study* (Zeitschrift für Erziehungswissenschaft, Sonderheft 14, pp. 283-299). Wiesbaden: VS Verlag.

American Association for Public Opinion Research (2015). *Standard Definitions Final Dispositions of Case Codes and Outcome Rates for Surveys.* 8th ed. Retrieved from https://www.aapor.org/AAPOR_Main/media/publications/StandardDefinitions2015_8theditionwithchanges_April2015_logo.pdf

Antoni, M., Drasch, K., Kleinert, C., Matthes, B., Ruland, M., & Trahms, A. (2010). *Arbeiten und Lernen im Wandel. Teil 1: Überblick über die Studie* (FDZ-Methodenreport 05/2010). Nürnberg: Bundesagentur für Arbeit.

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

Berlin M., Mohadjer L., Waksberg J., Kolstad. A., Kirsch I., Rock D., & Yamamoto K. (1992). An experiment in monetary incentives. In *Proceedings of the Survey Research Methods Section 1992* (pp. 393-398). Alexandria, VA: American Statistical Association.

Bethlehem, J., Cobben, F., & Schouten, B. (2011). *Handbook of nonresponse in household surveys.* New York: John Wiley & Sons.

Blohm, M., & Koch, A. (2013). Respondent incentives in a national face-to-face survey. Effects on outcome rates, sample composition and fieldwork efforts. *Methoden, Daten, Analysen, 7*(1), 89-122.

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

Blossfeld, H.-P., & von Maurice, J. (2011). Education as a lifelong process. In H.-P. Blossfeld, H.-G. Roßbach & J. von Maurice (Eds.), *Education as a lifelong process. The German National Educational Panel Study* (Zeitschrift für Erziehungswissenschaft, Sonderheft 14, pp. 19-34). Wiesbaden: VS Verlag.

Börsch-Supan, A., Krieger, U., & Schröder, M. (2013). *Respondent incentives, interviewer training and survey participation* (SHARE Working Paper Series 12-2013). Retrieved from http://www.shareproject.org/uploads/tx_sharepublications/WP_Series_12_2013_B%C3%B6rsch-Supan_Krieger_Schr%C3%B6der_02.pdf

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

Dillman, D.A., Smyth, J.D., & Christian, L.M. (2014). *Internet, phone, mail and mixed-mode surveys: The tailored design method. 4th edition*. Hoboken, NJ: John Wiley & Sons.

Gouldner, A. (1960). The norm of reciprocity: A preliminary statement. *American Sociological Review*, *25*(2), 161-178.

Groves, R. M., Couper, M. P., Presser, S., Singer, E., Tourangeau, R., Acosta, G. P., & Nelson, L. (2006). Experiments in producing nonresponse bias. *Public Opinion Quarterly*, *70*(5), 720-736.

Groves, R.M., Dillman, D. A., Eltinge, E. J. & Little, R. J. A. (2002). *Survey nonresponse*. New York: John Wiley & Sons.

Groves, R. M., Fowler, F. J., Couper, M. P., Lepkowski, J. M., Singer, E. & Tourangeau, R. (2004). *Survey methodology*. Hoboken, NJ: John Wiley & Sons.

Groves, R. M., & Peytcheva, E. (2008). The impact of nonresponse rates on nonresponse bias: A meta-analysis. *Public Opinion Quarterly*, *72*(2), 167-189.

Groves R. M., Singer, E. & Corning, A. (2000). Leverage-Salience Theory of Survey Participation: Description and an Illustration. *Public Opinion Quarterly*, 64(3), 299-308.

infas Institut für angewandte Sozialwissenschaft (2009). *Nationales Bildungspanel (NEPS) Etappe 8: Adult Education and Lifelong Learning. Erste Welle der Erwachsenenetappe* (Methodenbericht Machbarkeitsstudie Dezember 2009). Bonn: Institut für angewandte Sozialwissenschaft.

James, T. L. (1997). Results of the wave 1 incentive experiment in the 1996 Survey of Income and Program Participation. In *Proceedings of the Survey Research Methods Section 1997* (pp. 834-839). Alexandria, VA: American Statistical Association.

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

Kleinert, C., Christoph, B. & Ruland, M. (2015). Auswirkungen der Administration von Kompetenztests im Rahmen einer Panelerhebung für Erwachsene. Ergebnisse eines Experiments in Startkohorte 6 des Nationalen Bildungspanels (NEPS). In J. Schupp & C. Wolf (Eds.) *Nonresponse Bias. Qualitätssicherung sozialwissenschaftlicher Umfragen*. (pp. 359-382). Wiesbaden: Springer VS.

Laurie, H. (2007). *The effect of increasing financial incentives in a panel survey: An experiment on the British Household Panel Survey, Wave 14* (ISER Working Paper 2007-5). Colchester: University of Essex.

Laurie, H., & Lynn, P. (2009). The use of respondent incentives on longitudinal surveys. In P. Lynn (Ed.), *Methodology of longitudinal surveys* (pp. 205-233). Chichester: John Wiley & Sons.

Leibniz-Institut für Bildungsverläufe e.V. (2015). *Startkohorte 6: Erwachsene (SC6). Studienübersicht. Wellen 1 bis 5* (Documentation). Bamberg: Leibniz-Institut für Bildungsverläufe e.V.

Mack, S., Huggins, V., Keathley, D., & Sundukchi, M. (1998). Do monetary incentives improve response rates in the Survey of Income and Program Participation? In *Proceedings of the Survey Research Methods Section 1998* (pp. 529-534). Alexandria, VA: American Statistical Association.

Mann, S. L., Lynn, D. J., & Peterson A. V. (2008). The "downstream" effect of token prepaid cash incentives to parents on their Young Adult Children`s Survey participation. *Public Opinion Quarterly, 72*(3), 487-501.

Mercer, A., Caporaso, A., Cantor, D., & Townsend, R. (2015). How much gets you how much? Monetary incentives and response rates in household surveys. *Public Opinion Quarterly, 79*(1), 105-129.

Pew Research Center. (2012). *Assessing the representativeness of public opinion surveys.* Washington, DC: Pew Research Center. Retrieved from http://www.people-press.org/files/legacypdf/Assessing%20the%20Representativeness%20of%20Public%20Opinion%20Surveys.pdf

Pforr, K., Blohm, M., Blom, A. G., Erdel, B., Felderer, B., Fräßdorf, M., … Rammstedt, B. (2015). Are incentive effects on response rates and nonresponse bias in large-scale, face-to-face surveys generalizable to Germany? Evidence from ten experiments. *Public Opinion Quarterly*, 79(3), 740-768.

Pohl, S., & Carstensen, C. H. (2012). *NEPS technical report – Scaling the data of the competence tests* (NEPS Working Paper No. 14). Bamberg: Leibniz-Institut für Bildungsverläufe e.V.

Rodgers, W.L. (2002). Size of incentive effects in a longitudinal study. In *Proceedings of the Survey Research Methods Section 2002* (pp. 2930-2935). Alexandria, VA: American Statistical Association.

Rodgers, W.L. (2011). Effects of increasing the incentive size in a longitudinal study. *Journal of Official Statistics, 27*(2), 279-299.

Ryu, E., Couper, M. P., & Marans, R. W. (2005). Survey incentives: Cash vs. in-kind; face-to-face vs. mail; response rate vs. nonresponse error. *International Journal of Public Opinion Research*, *18*(1), 89–106.

Scherpenzeel, A., & Toepoel, V. (2012). Recruiting a probability sample for an online panel: Effects of contact mode, incentives, and information. *Public Opinion Quarterly, 76*(3), 470-490.

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

Schupp, J. & Wolf, C. (2015). *Nonresponse Bias. Qualitätssicherung sozialwissenschaftlicher Umfragen*. Wiesbaden: Springer VS.

Singer, E. (2002). The use of incentives to reduce nonresponse in household surveys. In R. M. Groves, D. A. Dillman, J. L. Eltinge & R. J. A. Little (Eds.), *Survey nonresponse* (pp. 163-178). New York: John Wiley & Sons.

Singer, E., van Hoewyk, J., Gebler, N., Raghunathan, T., & McGonagle, K. (1999). The effect of incentives on response rates in interviewer-mediated surveys. *Journal of Official Statistics, 15*(2), 217-230.

Willimack, D. K., Schumann, H., Pennell, B.-E., & Lepkowski, J. M. (1995). Effects of a prepaid nonmonetary incentive on response rates and response quality in a face-to-face survey. *Public Opinion Quarterly, 59*(1), 78-92.

Yu, J., & Cooper, H. (1983). A quantitative review of research design effects on response rates to questionnaires. *Journal of Marketing Research, 20*(1), 36-44.

Zagorsky, J. L., & Rhoton, P. (2008). The effects of promised monetary incentives on attrition in a long-term panel survey. *Public Opinion Quarterly, 72*(3), 502-513.

Zinn, S., Aßmann, C., & Würbach, A. (2015). *Sampling and weighting the sample of the adult cohort of the National Educational Panel Study (Wave 2 to 5). Technical Report on SUF SC6, Version 5.0.0* (Technical Report). Bamberg: Leibniz-Institute for Educational Trajectories.

Appendix

Table A1

	All		postpaid				(partly) prepaid				Interview: col (2) - (1)	
	%	n	%	n	Interview %	n	%	n	Interview %	n	%	p-value
Sex												
Male	49.8	6,110	49.2	3,026	77.5	2,345	50.3	3,084	79.1	2,439	1.6	0.128
Female	50.2	6,170	50.8	3,120	78.6	2,453	49.7	3,050	81.2	2,476	2.6	0.019
valid	*100.00*	*12,280*	100.00	*6,146*			100.00	*6,134*				
Age												
>30	3.4	417	3.6	219	66.2	145	3.2	198	69.7	138	3.5	0.440
30-39 years	15.9	1,951	15.8	971	72.6	705	16.0	980	72.0	706	-0.6	0.774
40-49 years	24.8	3,042	24.9	1,532	77.3	1,185	24.6	1,510	79.9	1,207	2.6	0.071
50-59 years	33.1	4,068	33.3	2,045	80.7	1,651	33.0	2,023	82.6	1,670	1.8	0.133
60+ years	22.8	2,800	22.4	1,378	80.6	1,111	23.2	1,422	83.9	1,193	3.3	0.024
valid	*100.00*	*12,278*	100.00	*6,145*			100.00	*6,133*				
Migration background												
No	84.7	10,401	84.8	5,209	79.2	4,124	84.6	5,192	81.2	4,214	2.0	0.012
Yes	15.3	1,879	15.2	937	71.9	674	15.4	942	74.4	701	2.5	0.232
valid	*100.00*	*12,280*	100.00	*6,146*			100.00	*6,134*				

Table A1 continued

	All		postpaid				(partly) prepaid				Interview: col (2) - (1)	
	%	n	%	n	Interview %	n	%	n	Interview %	n	%	p-value
Employment status												
(Self-)employed	76.8	9,143	77.0	4,591	80.3	3,686	76.5	4,552	82.1	3,739	1.9	0.027
Unemployed	4.0	478	3.9	234	79.5	186	4.1	244	75.0	183	-4.5	0.213
Retired	12.2	1,449	11.9	708	85.6	606	12.5	741	87.9	651	2.3	0.204
Family care	4.5	531	4.6	273	76.2	208	4.3	258	86.0	222	9.9	0.004
Education/civil service	1.1	128	1.0	60	61.8	37	1.1	68	61.8	42	0.1	0.991
Other	1.5	183	1.6	95	78.9	75	1.5	88	87.5	77	8.6	0.138
valid	*100.00*	*11,912*	*100.00*	*5,961*			*100.00*	*5,951*				
Marital status												
Single	16.1	1,974	16.2	997	77.1	769	15.9	977	77.8	760	0.7	0.748
Married, living together	64.7	7,944	64.6	3,967	79.3	3,144	64.8	3,977	81.7	3,251	2.5	0.007
Married, living apart	2.1	257	2.2	132	81.8	108	2.0	125	84.8	106	3.0	0.534
Partner, living together	11.2	1,373	11.2	690	74.6	515	11.1	683	78.0	533	3.4	0.124
Partner, living apart	6.0	732	5.9	360	72.8	262	6.1	372	71.2	265	-1.5	0.643
valid	*100.00*	*12,280*	*100.00*	*6,146*			*100.00*	*6,134*				
Children in household												
None	49.0	4,179	48.3	2,050	86.8	3,684	49.7	2,129	89.0	3,810	2.3	0.002
0 to 3 years old	7.6	650	7.3	308	88.6	273	8.0	342	88.3	302	-0.3	0.892
4 to 15 years old	30.2	2,576	30.5	1,295	87.0	1,127	29.9	1,281	86.9	1,113	-0.1	0.919
16+ years old	33.2	2,827	33.6	1,426	86.0	1,227	32.6	1,401	87.4	1,224	1.3	0.318

Table A1 continued

	All		postpaid				(partly) prepaid				Interview: col (2) - (1)	
	%	n	%	n	Interview %	n	%	n	Interview %	n	%	p-value
Household income												
Less than 1,500	11.1	1,291	11.1	652	74.7	487	11.0	639	72.9	466	-1.8	0.472
1,500 to 3,000	33.8	3,943	33.8	1,976	76.4	1,509	33.9	1,967	80.2	1,577	3.8	0.007
More than 3,000	55.1	6,420	55.1	3,225	79.8	2,574	55.1	3,195	82.0	2,620	2.2	0.016
valid	*100.00*	*11,654*	*100.00*	*5,853*			*100.00*	*5,801*				
BIK												
Up to 50,000	24.2	2,975	24.7	1,516	79.4	1,203	23.8	1,459	80.7	1,178	1.4	0.424
50,000 to 100,000	10.9	1,340	10.6	654	76.6	501	11.2	686	79.6	546	3.0	0.193
100,000 to 500,000	32.4	3,975	32.6	2,001	77.3	1,547	32.2	1,974	80.4	1,588	3.1	0.011
More than 500,000	32.5	3,990	32.1	1,975	78.3	1,547	32.9	2,015	79.6	1,603	1.2	0.348
valid	*100.00*	*12,280*	*100.00*	*6,146*			*100.00*	*6,134*				

Table A2 Wave 6 Response Rates by Mathematics Proficiency (N= 5,645)

	postpaid			(partly) prepaid				
		Interview			Interview			
	(n= 2,811)	%		(n= 2,834)	%		col (2) - (1)	p-value
No test	600	72.0	*432*	633	75.2	*476*	3.2%	0.154
Lower tercile	747	82.6	*617*	723	84.9	*614*	2.3%	0.225
Middle tercile	709	83.4	*591*	765	84.7	*648*	1.3%	0.505
Higher tercile	755	87.0	*657*	713	87.5	*624*	0.5%	0.784

Can Interviewer Observations of the Interview Predict Future Response?

Ian Plewis[1], *Lisa Calderwood*[2] & *Tarek Mostafa*[2]

1 *University of Manchester*

2 *Centre for Longitudinal Studies, UCL Institute of Education*

Abstract

Interviewers made four observations related to future participation, respondent cooperation, enjoyment and whether the respondent found the questions difficult, for a large sample of face-to-face interviews at wave four of the UK Millennium Cohort Study (MCS). The focus of the paper is on predicting response behavior in the subsequent wave of MCS, four years later. The two most predictive observations are whether the respondent is likely to participate in the next wave and whether they enjoyed the interview. Not only do these predict non-response at the next wave, they do so after controlling for other explanatory variables from earlier waves in a response propensity model. Consequently, these two interviewer observations improve discrimination between respondents and non-respondents at wave five as estimated by Gini coefficients generated by a Receiver Operating Characteristic curve analysis. The predicted probabilities of responding at wave five are also used to estimate R-indicators, particularly to address the question of whether, hypothetically, conversion of 'frail' respondents would lead to improved representativity and reduced bias in longitudinal estimates of interest. The evidence from the R-indicators and partial R-indicators suggests that successful conversions could achieve those aims although the cost of so doing might outweigh the benefits.

Keywords: Millennium Cohort Study; non-response; representativity; response propensity models; ROC curve.

1 Introduction

An important goal for managers of longitudinal surveys is to maintain response over time so that researchers can have some confidence in their inferences about change. Various strategies are used: incentives (both to respondents and interviewers), reissuing refusals etc. Many of these issues are discussed in Lynn (2009). Another possibility is to direct extra resources at those respondents with a higher risk of not responding, a risk that is often estimated from response propensity models that include predictors from previous waves. Often, however, predictions of future non-response are imprecise so that targeted interventions might not be cost-effective (Plewis & Shlomo, 2013). Our paper focuses on interviewer observations of a face-to-face interview. We investigate the characteristics of these observations and whether they can improve the prediction of non-response at the subsequent wave of data collection, both on their own and, more importantly, over and above the variables that are commonly included in response propensity models. We then go on to consider the implications for the longitudinal sample of a hypothetical situation in which respondents deemed to be at high risk of not responding at the subsequent wave are retained in the sample.

Interest in the value of collecting interviewer observations of the characteristics of neighborhoods, the quality and type of dwelling units and the circumstances of respondents has expanded in recent years as part of a more general interest in survey paradata (Kreuter, 2013). To the extent that interviewer observations of this kind are correlated both with the propensity to respond and with survey variables of interest, they might profitably be used to reduce bias arising from non-response as discussed, in a cross-sectional context, by Kreuter et al. (2010). Interviewer observations of their own interviews – the focus of this paper - have attracted less

Acknowledgements

The interviewer observations reported in this paper were collected as part of a research project entitled ‘Predicting and Preventing Non-Response in Cohort Studies’ (Ref: RES-175-25-0010). This project was funded by the UK’s Economic and Social Research Council (ESRC) as part of their Survey Design and Measurement Initiative (SDMI). The UK Millennium Cohort Study (MCS) is funded by the ESRC and a consortium of UK government departments. The National Centre for Social Research (NatCen) carried out the data collection in Great Britain for the fourth wave of MCS; we would like to thank their research and operational staff who were responsible for the implementation of the interviewer observations. Ipsos MORI carried out the data collection for the fifth wave of MCS. We would also like to thank Rebecca Taylor for her contributions to this research, and the referees and Editor for helpful comments on an earlier version of this manuscript.

Direct correspondence to

Ian Plewis, Social Statistics, Humanities Bridgeford Street,
Manchester M13 9PL, UK
E-mail: ian.plewis@manchester.ac.uk

attention from researchers. Eckman et al. (2013) provide a summary although none of the studies reviewed by them are in peer-reviewed journals. The context for the empirical investigation in Eckman et al. (and also in Sinibaldi & Eckman, 2015) is a German cross-sectional telephone survey. Essentially, interviewers were asked to rate the probability that the case would complete the interview at a later contact attempt (conditional on them not doing the interview at that contact). The authors do find that the higher the probability rating the more likely a subsequent interview, although the association appears to be non-linear and not to be strong. Sinibaldi & Eckman (2015) extend the analysis by showing that discrimination between completion and non-completion is slightly improved when the interviewer ratings are added to a response propensity model that already includes other 'call' variables to predict outcome. They also consider how these ratings might be used in a hypothetical adaptive design to improve cooperation rates. Neither Eckman et al. nor Sinibaldi & Eckman address the question of whether these interviewer variables will lead to a reduction of non-response bias in outcomes of interest.

Few studies have used interviewer observations in a longitudinal context. We have previously shown (Plewis et al., 2012) that interviewer observations of neighborhood at wave two in the study used in this paper - the ongoing UK birth cohort study known as the Millennium Cohort Study (MCS) - predict response one wave later. West et al. (2014) collected interviewer ratings of income (in terciles) and whether the respondent was receiving unemployment benefit to supplement survey measures of these variables. They found that, in terms of non-response adjustment, these observations do not have any additional effect on their chosen cross-sectional estimates having incorporated prior survey measures of economic variables in their response propensity model. Uhrig (2008), using data from waves one to 14 of the British Household Panel Survey, shows that an interviewer rating at the end of the interview of respondent cooperativeness during the interview (a five point scale) predicts later response, after controlling for other variables in a discrete time hazard model with attrition as an absorbing state. He modeled non-contact (a category that includes not located) and refusal separately and found that the model estimates increase monotonically across the five point scale and are statistically significant for both response categories although they are stronger for refusal. None of this cited work considers how interviewer observations might be used in adaptive longitudinal designs to maintain response over time.

Our paper builds on this rather small body of research. We consider whether previous findings on associations with non-response, and on discrimination between respondents and non-respondents, are replicated with a broader set of interviewer observations of the interview process. We also consider the potential value of these ratings for improving estimates of the representativity of longitudinal samples at wave *t+1* in terms of the wave *t* sample, and for targeting interventions at what we call 'frail' respondents in the context of a hypothetical adaptive design.

The paper is organized as follows. Section 2 describes the data used for our empirical investigations and presents some basic descriptive statistics. Section 3 sets out our research questions in their statistical modeling context. Section 4 presents the results from our models. Section 5 concludes with some reflections on our results and their implications for future longitudinal investigations.

2 Data

The data for this investigation come from a methodological study incorporated into wave four of the UK Millennium Cohort Study (MCS). Wave one of MCS includes children from 18,552 families born over a 12-month period during the years 2000 and 2001, and living in selected UK electoral wards at age nine months. The initial response rate was 72%. Areas with high proportions of Black and Asian families, disadvantaged areas and the three smaller UK countries are all over-represented in the sample which is disproportionately stratified and clustered as described in Plewis (2007). The first five waves took place when the cohort members were (approximately) nine months, 3, 5, 7 and 11 years old (in 2012). The data collection for the study takes place in the home and involves face-to-face interviews with multiple informants in each family. Interviews have been sought with up to two co-resident parents at every wave. At wave five, 31% of the target sample – which excludes child deaths and emigrants – were unproductive in the sense of not providing any data (Mostafa, 2014).

During wave four of MCS, interviewers were asked to rate (using five point scales) some aspects of the interview after it was completed: whether participation was likely at the next sweep (i.e. wave); and observations of (i) cooperation during the interview and (ii) whether the respondent had enjoyed the interview. In addition, interviewers were asked to assess whether the respondents had found answering any of the questions difficult or uncomfortable. The motivation for the first three of these observations is clear in terms of the previously cited literature and their face validity; the final observation was included because it was expected to tap an aspect of the interview more closely related to the actual interaction between interviewer and respondents. Appendix A gives the wording for the interviewers when making the observations.

In principle, both main respondents (usually mothers of the cohort child) and their partners (if present in the household) answered survey questions. Hence, all observations apart from the one about likely future participation were recorded by the interviewers for both respondents and partners. There was a tendency for main respondents to be given more positive ratings than their partners, and also for main respondents with partners who responded to be rated more positively than main respondents as a whole. The exception was the ‘questions difficult’ obser-

Table 1 Percentage distributions of interviewer observations

	SCALE VALUE [1]					
OBSERVATION	1	2	3	4	5	n
Future participation	82	15	2	*	*	13099
Enjoyment [2]	39	47	13	1	*	13059
Cooperation [2]	73	23	4	*	*	13058
Questions difficult [2]	89	11	n.a.	n.a.	n.a.	12811 [3]

Notes:

[1] Scale value '1' represents the positive end of the scale, '3' is neutral ('difficult to say' or 'fair'), '5' the most negative. *: < 0.5%.

[2] Main respondent and partner observations were combined in such a way that the more negative rating was dominant. When there was no partner interview, the main respondent rating was used (and vice-versa).

[3] 2% of respondents who were rated as 'not sure/don't know' are omitted.

vation where responding partners were perceived to have found the questions, if anything, less difficult and uncomfortable. Agreement between the observations for main respondents and their partners (aggregated over interviewers) was moderate: the kappa estimates (weighted to reflect the extent of disagreement) are 0.50 (s.e. = 0.01; n = 8739) for enjoyment; 0.43 (0.01; 8741) for cooperation and 0.40 (0.01; 8741) for the binary 'questions difficult'. We do not know whether decisions about participating in MCS are made independently or jointly within households. In this paper, we concentrate on predicting non-response at the household level, treating as responding any household that provides at least some data. Consequently, we combine the respondent and partner assessments to generate a single variable for modeling response propensities and we do this by taking the more negative rating for each observation if two observations were made. This does assume that decisions are more likely to be made jointly by the main respondent and her/his partner and has the advantage, in the modeling, of having variables which are less skewed to the positive end of the scale and show more variation.

Table 1 gives the descriptive results for the four interviewer observations. It shows that all four are skewed towards the positive ends of the scales although less so for 'enjoyment'. The participation, enjoyment and cooperation questions all correlate moderately with each other but there is no correlation between 'questions difficult' and the other three variables which suggests that this observation is, as anticipated, tapping a different dimension of the interview. As our main interest is in analyzing response at wave five, we treat the issued sample at wave five that was productive at wave four (n = 13108) as our base sample. Overall non-response is 11%. Most of the non-response comes from cases who refuse (n = 1102; 8% of all cases); not located (i.e. untraced) is 1.1% (n = 155) and non-contact conditional on being traced is 1.7% (n = 218). There was very little non-response – less than 1% - for the interviewer observations as indicated by the final column of Table 1. The percentages in Table 1 allow for the sample design (disproportionate stratification and clustering); sample sizes (n) are the actual number of observations.

The child's ethnic group and the highest level of educational qualifications achieved by the main respondent are key socio-demographic variables in MCS in that they are associated with many of the economic, social, health and cognitive outcomes of interest. We therefore assess whether these key variables are associated with the interviewer observations. We find that, when these variables are explanatory variables in ordered (i.e. proportional odds) and binary logistic regressions, they both predict all the interviewer observations. Interviewers expect participation at the next wave to be less likely among the mixed, Pakistani and Bangladeshi, and Black Caribbean and African ethnic groups than for whites, Indians and others; $p < 0.001$ on a Wald test. The results for enjoyment, cooperation and 'questions difficult' are similar although not identical. Pakistani and Bangladeshi, Black Caribbean and African, and 'other' ethnic groups are assessed to have enjoyed the interview less and to have been less cooperative whereas all the minority ethnic groups apart from the mixed group were more likely to have found the questions difficult (Wald tests all $p < 0.001$). Mothers with lower qualifications were more likely to be assessed at the more negative points on all four scales (Wald tests all $p < 0.001$).

3 Methods and Models

We fit statistical models to answer three questions. The first is whether interviewer observations at wave *t* predict overall non-response, and categories of non-response, at wave *t+1*, both separately and when put together in a single model. Moreover, do these observations predict response at wave *t+1* conditional on the inclusion in a response propensity model of established explanatory variables from previous waves? The full response propensity model is:

$$logit\left(\rho_i\right) = \sum_{k=0}^{K}\beta_k x_{ki} + \sum_{l=1}^{L}\gamma_l z_{li} \tag{1}$$

where $\rho_i = E(r_i)$ is the probability of responding for unit i ($i = 1..n$); $r_i = 0$ for non-response and 1 for response; x_k are the explanatory variables from previous waves and listed in Appendix B ($x_0 = 1$); z_l are the interviewer observations. ML estimates of β_k (= b_k) and γ_l (= c_l) are easily obtained, leading to predicted probabilities or propensities of responding $\hat{\rho}_i$ where

$$\hat{\rho}_i = e^{\sum_{k=0}^{K} b_k x_{ki} + \sum_{l=1}^{L} c_l z_{li}} \Big/ \left(1 + e^{\sum_{k=0}^{K} b_k x_{ki} + \sum_{l=1}^{L} c_l z_{li}} \right) \tag{2}$$

The second question is: how much improvement is provided by the interviewer observations in terms of discriminating between respondents and non-respondents at wave *t+1*, as measured by analyses using Receiver Operating Characteristic (ROC) curves? Our approach to this question is based on estimating the predicted probabilities ($\hat{\rho}_i$) of responding at wave five from the response propensity models without and with interviewer assessments. It is set out in detail in Plewis et al. (2012). We present just the essentials of this method here.

Plewis et al. (2012) show how ROC curves can be used to discriminate between, or to predict whether cases are more likely to be respondents or non-respondents. In brief, if + (i.e. $1 - \hat{\rho}_i > c$) refers to a prediction of non-response where *c* is any threshold from the distribution of $\hat{\rho}_i$ then the ROC is the plot of $P(+| r = 0)$ against $P(+| r = 1)$ where *r* is the observed response category, i.e. a plot of the true positive fraction (TPF) against the false positive fraction (FPF) for all *c*.

The area enclosed by the ROC curve and the x-axis, known as the AUC (area under the curve), is of particular interest and this can vary from 1 (when the model for predicting response perfectly discriminates between respondents and non-respondents) down to 0.5, the area below the diagonal (when there is no discrimination between the two categories). The AUC can be interpreted as the probability of assigning a pair of cases, one respondent and one non-respondent, to their correct categories, bearing in mind that guessing would correspond to a probability of 0.5. A linear transformation of AUC (= 2*AUC – 1), often referred to as a Gini coefficient, is commonly used as a more natural measure than AUC because it varies from 0 to 1.

Plewis et al. (2012) also use a method developed by Copas (1999) known as a logit rank plot. For response propensity models based on logistic regression, this is just a plot of the linear predictor from the model against the logistic transformation of the proportional rank of the propensity scores. Copas argues that this approach is more sensitive to changes in the response propensity model than an approach based on ROC curves.

The third question is: what are the implications for the characteristics of the longitudinal sample of (i) using the interviewer observations in a response propensity model and (ii) hypothetically converting to respondents those non-respondents at wave *t+1* who were observed by the interviewers to be 'frail' respondents at wave *t*? We use R-indicators to answer the two parts of this question. The R-indicator is described by Schouten et al. (2009); in essence, it is an overall measure of how far the observed sample deviates from the target sample in terms of likely bias. It is estimated by:

$$\hat{R}_{\rho} = 1 - 2\hat{S}_{\rho} \tag{3}$$

where ρ is the probability of responding, estimated from the response propensity model as in (2), and $\hat{S}_{\rho}$ is the standard deviation of these estimated probabilities. Standard errors of $\hat{R}_{\rho}$ for clustered and weighted samples are discussed by Plewis & Shlomo (2013). It is important to note that the estimate of *R* is conditional on the specification of the response propensity model.

We also use unconditional partial R-indicators ($R_{p(u)}$) for the third question. Unconditional partial R-indicators for a variable *Z* having categories *j, j = 1..J* show how representativeness varies across this variable and thus provides an indication of where the sample is particularly deficient (or satisfactory). Conditional on the response propensity model, the variable level unconditional partial R-indicator is estimated as:

$$\hat{R}_{p(u)} = \sqrt{[\sum_{j=1}^{J} p_j (\hat{\rho}_j - \hat{\rho})^2]}$$

where p_j is the estimated proportion in category *j*, $\hat{\rho}_j$ is the estimated (mean) response rate in category *j* and $\hat{\rho}$ is the estimated overall response rate. A reduction in $\hat{R}_{p(u)}$ indicates an improvement in representativeness with respect to that variable.

At the category level, *Z = j*, the unconditional partial indicator is estimated as:

$$\hat{R}_{p(u),j} = \sqrt{p_j}\ (\hat{\rho}_j - \hat{\rho})$$

Note that $\hat{R}_{p(u),j}$ can be negative (under-representation) or positive (over-representation).

4 Results

Here, we give the results for the three questions posed in the previous section.

4.1 Are interviewer observations predictive?

All four interviewer observations from wave four (i.e. *t*) predict overall non-response at wave five (*t+1*) as shown by the estimates from the logistic regressions in Table 2. The estimates increase monotonically except for the final categories which have few observations (see Table 1).

Table 2 Estimates from logistic regressions for each observation

	Estimate (s.e.)				
OBSERVATION	2	3	4	5	n
Future participation	-1.02 (0.089)	-1.71 (0.16)	-2.52 (0.37)	-1.43 (0.42)	13099
Enjoyment	-0.44 (0.084)	-0.94 (0.11)	-1.59 (0.22)	-1.27 (0.35)	13059
Cooperation	-0.56 (0.073)	-1.14 (0.13)	-1.35 (0.35)	-1.37 (0.53)	13058
Questions difficult	-0.52 (0.094)	n.a.	n.a.	n.a.	12811

Notes

1. The reference category is the most positive rating.
2. The models are fitted using the *svy* procedures in STATA and so allow for the sample design.

When wave five non-response is broken down into not located, not contacted and refusal, we find that ‘future participation’ and ‘questions difficult’ predict all three non-response categories but ‘enjoyment’ and ‘cooperation’ only predict refusal (conditional on being contacted) and non-contact (conditional on being located). The fact that the observation of likely future participation predicts whether or not someone is located at the next wave suggests that interviewers pick up clues during or after the interview about family plans to move, making it difficult to interpret this association. Because non-contacts are sometimes regarded as disguised refusals (Blom, 2014), and because the relations between the observations and these two categories are similar, we combine these two categories and omit the not located cases from the rest of the analyses presented here. Hence, we work with a new binary variable r^*: refused or not contacted ($r^* = 0$) and responded (or productive) ($r^* = 1$).

When all four interviewer observations are entered together into a single model, we find that ‘future participation’ and ‘enjoyment’ conditionally predict r^* but ‘cooperation’ and ‘questions difficult’ do not. The estimates and p-values from Wald tests from the logistic regression model are: (-0.84, -1.39, -2.25, -1.44), $p<0.001$ (‘future participation’); (-0.21, -0.41, -0.56, -0.28), $p<0.03$ (‘enjoyment’); (-0.013, -0.10, -0.065, -0.40), $p>0.9$ (‘cooperation’); 0.16, $p >0.15$ (‘questions difficult’). Consequently, we focus on ‘future participation’ and ‘enjoyment’ from now on.

We do find that both ‘future participation’ and ‘enjoyment’ predict r^* after controlling for all other variables. The estimates for these two observations are given in Table 3 (and the full set of estimates is given in Appendix B). In other words, interviewer observations can improve the prediction of non-response beyond what can be achieved with the usual response propensity models in longitudinal research. The extent of that improvement is now addressed.

Table 3 Estimates for the two interviewer observations in the full response propensity model

OBSERVATION	Estimate (s.e.)				
	2	3	4	5	n
Future participation	-0.58 (0.11)	-0.94 (0.19)	-1.97 (0.41)	-1.45 (0.44)	12880
Enjoyment	-0.28 (0.090)	-0.50 (0.13)	-0.82 (0.25)	0.39 (0.41)	

4.2 Is discrimination improved?

The two interviewer observations increase the AUC from 0.68 (s.e. = 0.0079) to 0.70 (s.e. = 0.0076). This difference is greater than expected by chance ($\chi_1^2 = 23.8, p < 0.001; n = 12880$) from the roccomp procedure in STATA. This means the Gini coefficient increases from 0.36 to 0.41. The slopes of the logit rank plots tell a similar story: an increase from 0.38 (0.011) to 0.43 (0.013).

These results indicate that the two more predictive interviewer observations do improve the prediction of non-response. Whether this model would also be better for adjusting for non-response using non-response weights or imputation methods, does, however, require that the observations are correlated with outcome variables of interest, more particularly changes in these variables, as well as with response behavior. This is also one of the requirements for targeting interventions at potential non-respondents although maintaining the sample over time does also have benefits in terms of precision. We do not address this question directly here but return to it in the concluding section.

4.3 Implications for representativity?

We find that the response propensity model that includes the two interviewer observations leads to a reduced estimate of *R* (0.83) compared with the model without them (0.86). Using the methods described in Plewis & Shlomo (2013), this difference is greater than would be expected by chance. In other words, the improved response propensity model not only discriminates better between respondents and non-respondents (as shown by the Gini coefficients etc.), it also provides a lower and what is probably a better estimate of how representative the wave five sample is in terms of the productive sample at wave four.

Given that the interviewer observations at wave *t* are predictive of response at wave *t+1* and taking advantage of the fact that they can be made available to survey managers soon after fieldwork for wave *t* has been completed, another way of using them is to define a set of what we might call 'frail' respondents who have a low rating (i.e. 3 or below) on at least one of the two most predictive observations. In principle, it would be possible to direct extra resources (such as using more experienced interviewers or financial incentives) at these 'frail' respondents with the intention of preventing them from becoming non-respondents at the next wave.

There were 352 frail respondents as defined above who were indeed non-respondents at wave five. We use the response propensity model without the interviewer observations to estimate *R*. Were our interventions to convert all these non-respondents into respondents at wave five successful, then the estimate of *R* would increase from 0.86 (the estimate given above) to 0.91. Of course, no interventions to prevent non-response will have a 100% success rate. Moreover, any intervention will also be directed at 'frail' respondents who did, in the event, respond at wave five: there were 1838 of these in our example so the targets of the intervention would form perhaps only a sixth of the intervention group. We could reduce this 'deadweight' problem by having a stricter criterion such as respondents receiving a rating in just the two lowest categories for at least one of the observations. This would reduce the size of the intervention group to 290 of which 63 (22%) actually failed to respond at wave five. The effect on representativity is then smaller (0.87 compared with 0.86). Nevertheless, this approach does demonstrate the possibilities of combining interviewer observations with targeted interventions in terms of maintaining the sample over time and reducing the overall bias in the sample. We can provide at least some evidence about whether non-response bias in outcome variables of interest will be reduced by estimating unconditional partial R-indicators for the two key variables introduced earlier – ethnic group and qualifications.

We find that the unconditional partial R-indicator for ethnic group would decline slightly - from 0.018 to 0.014 - if frail respondents were maintained in the sample (using the less strict criterion of frailty). The decline in $\hat{R}_{p(u)}$ for qualifications is more marked: 0.031 to 0.021. The estimates of $\hat{R}_{p(u),j}$ show that under-representation of the mixed and black groups, and the over-representation of the highly qualified groups, would both be reduced. This suggests that keeping the frail respondents in the sample might lead to a reduction in bias in estimates of interest.

5 Discussion

We have shown that interviewers are willing and able to make observations of their interviews. It is, however, likely that interviewers vary in the way they generate observations of this kind. Eckman et al. (2013) show that, in their study with

34 interviewers randomly assigned to cases in their telephone survey, about nine per cent of the variation in their one rating could be attributed to interviewers. About 400 interviewers were used in wave four of MCS and, as is common in such large face-to-face longitudinal surveys, they were not randomly allocated to cases. Consequently, we have no estimate of the interviewer effect for our observations although we can be sure that interviewers will have observed 'similar' interviews in different ways. It is probable that the variation between interviewers, if estimable, would have had a small effect on the estimates in our models, the most likely effect being to increase their standard errors. If the proportion of overall variation allocated to interviewers for our observations were similar to the estimate found by Eckman et al. (2013), and given a mean interviewer workload of about 30 cases, then we might expect to see a doubling of the standard errors. Most of our results are robust to such a reduction in the estimates' precision. Further investigation of this topic is, however, warranted.

This study used four interviewer observations; the only closely related study (Uhrig, 2008) used just one – a measure of cooperativeness – which did predict future response one year later. The evidence presented here suggests that an observation of cooperativeness is not as predictive as the observations of future participation and enjoyment. Hence, it is these two variables that researchers might consider giving priority to if they are in a position to collect such paradata in order to improve predictions via a better response propensity model. The 'questions difficult' variable does appear to be tapping another aspect of the interaction between interviewer and respondent but is not as good a predictor of future response as the others.

We have not directly addressed the question of whether the inclusion of interviewer observations into a response propensity model will reduce non-response bias in outcomes of interest. But we have shown that the observations are associated with key socio-demographic variables likely to be associated with changes in outcomes and so there are grounds for supposing that non-response weights based on the extended response propensity model will be more effective. Moreover, representativity in terms of these key variables is improved in our hypothetical adaptive design. Weighting is one way of trying to reduce non-response bias but it is not, of course, the only way. We can, for example, use multiple imputation in situations where, in our model of interest, we might have some unobserved outcomes (y) and explanatory variables (x) arising from item non-response and not from the unit non-response/attrition that weighting is designed to deal with. Interviewer observations might be useful in this context to predict both the missing y and the missing x. And, if the usual assumption of data missing at random (MAR) does not hold, we might want to use a Heckman selection model to adjust for non-response, jointly modelling the propensity to respond and the outcome of interest and allowing the residuals to be correlated. We then need instruments – variables associated with the propensity to respond and not with the outcome – for the model to be identified and interviewer observations measuring aspects of the interview itself could be useful instruments in that context.

We have focused here on the relation between interviewer observations and later non-response. It is, however, possible that observations of this kind could be used in other ways. In particular, they might be useful as accuracy indicators (Da Silva & Skinner, 2013) in order to get a handle on the extent of measurement error in the responses. It is plausible that the 'questions difficult' observation would be the most useful for this purpose. This is also a topic worthy of further investigation.

It remains an open question as to whether the benefits of collecting these kinds of interviewer observations outweigh their costs. Interviewers do have to be paid to complete these observations, perhaps only a small amount per interview, but a considerable sum in the aggregate. Hence, if field work budgets are fixed, some questions might, for example, have to be dropped from the questionnaire to accommodate them. The assessment of the benefits hinges on two related questions. First, would the incorporation of interviewer observations into a response propensity model lead to sufficiently improved non-response weights and imputations (i.e. greater bias reduction and more precision)? Second, would the retention of frail respondents in the sample as a result of a targeted intervention reduce bias and increase precision. This paper, along with Sinibaldi & Eckman (2015), does provide grounds for supposing that the answer to the first question could be positive. Both papers found, for example, similar increases (0.03 to 0.05) in the estimated Gini coefficients as a result of including observations in a response propensity model. The contexts for the two studies were, however, very different: a cross-sectional telephone survey with a low response rate and with predictions limited to a window of at most a few weeks, compared with an ongoing longitudinal study with high wave on wave response rates and predictions of response behavior four years later. An affirmative answer to the second question does depend on designing a successful intervention and being prepared to carry the cost of directing this intervention to a substantial 'deadweight' group of frail respondents who would have responded anyway.

Although this paper has a very specific focus on improving predictions of non-response, it can be located within the more general topic of assessing the value of paradata in longitudinal survey research. Combined with other research in this area, we are beginning to see a picture of how useful paradata might be in improving the quality of longitudinal data.

References

Blom, A. G. (2014). Setting priorities: Spurious differences in response rates. *International Journal of Public Opinion Research, 26(2)*, 245-255. doi:10.1093/ijpor/edt023

Copas, J. (1999). The effectiveness of risk scores: The logit rank plot. *Applied Statistics, 48(2)*, 165-183. doi:10.1111/1467-9876.00147

Da Silva, D. N., & Skinner, C. (2013). The use of accuracy indicators to correct for survey measurement error. *Applied Statistics, 63(2)*, 303-319. doi:10.1111/rssc.12022

Eckman, S., Sinibaldi, J., & Möntmann-Hertz, A. (2013). Can interviewers effectively rate the likelihood of cases to cooperate? *Public Opinion Quarterly*, 77(2), 561-573. doi:10.1093/poq/nft012

Kreuter, F. (Ed.) (2013). *Improving surveys with paradata: Analytic uses of process information*. Chichester: John Wiley & Sons.

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

Lynn. P. (Ed.) (2009). *Methodology of longitudinal surveys*. Chichester: John Wiley & Sons.

Mostafa, T. (2014). *Millennium Cohort Study: Technical Report on Response in Sweep 5*. Centre for Longitudinal Studies: London

Plewis, I. (2007). *The Millennium Cohort Study: Technical Report on Sampling (4th. Ed.)*. Centre for Longitudinal Studies: London.

Plewis, I., Ketende, S. C., Joshi, H., & Hughes, G. (2008). The contribution of residential mobility to sample loss in a birth cohort study: Evidence from the first two waves of the Millennium Cohort Study. *Journal of Official Statistics, 24(3)*, 365-385.

Plewis, I., Ketende, S., & Calderwood, L. (2012). Assessing the accuracy of response propensity models in longitudinal studies. *Survey Methodology, 38(2)*, 167-171.

Plewis, I., & Shlomo, N. (2013). Statistical guidance on optimal strategies to reduce non-response in longitudinal studies. *CCSR Working Paper 2013-03*. Manchester: CCSR.

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

Sinibaldi, J., & Eckman, S. (2015). Using call-level interviewer observations to improve response propensity models. *Public Opinion Quarterly, 79(4)*, 976-993. doi:10.1093/poq/nfv035

Uhrig, N. (2008). The nature and causes of attrition in the British Household Panel Survey. *Working Paper 2008-05*. ISER: Essex.

West, B. T., Kreuter, F., & Trappmann, M. (2014). Is the collection of interviewer observations worthwhile in an economic panel study? New evidence from the German labor market and social security (PASS) study. *Journal of Survey Statistics and Methodology, 2*, 159-181. doi:10.1093/jssam/smu002

Appendix A

Interviewer observations

This is how the four interviewer observations were worded:

1. In your opinion, how likely is it that anyone will take part in the next sweep of Child of the New Century: (1) very likely; (2) fairly likely; (3) difficult to say; (4) fairly unlikely; (5) very unlikely.
   [Child of the New Century is a label used by field staff to describe the Millennium Cohort Study.]

2. In general, how would you rate the co-operation of {main respondent (name)/ partner respondent (name)} during the interview: (1) very good; (2) good; (3) fair; (4) poor; (5) very poor.

3. On the whole, did {main respondent (name)/partner respondent (name)} seem to enjoy the interview: (1) enjoyed a great deal; (2) enjoyed to some extent; (3) difficult to say; (4) did not enjoy some of it; (5) did not enjoy at all.

4. During the interview did {main respondent (name)/partner respondent (name)} ever (a) seem to find the questions difficult, (b) indicate that it was taking a long time or (c) look uncomfortable when asked questions: yes to any; none of these; not sure/don't know.

Appendix B

Model estimates from response propensity model (in 4.1)

VARIABLE		ESTIMATE (s.e.)	95% CI
Child sex (ref: boy)		-0.21 (0.072)	(-0.35, -0.069)
Main respondent's age	<20	-0.18 (0.10)	(-0.37, -0.018)
(ref: 20-29)	30-39	0.36 (0.10)	(0.15, 0.56)
	40+	0.60 (0.44)	(-0.28, 1.47)
Ethnic group (ref: white)	Mixed	-0.15 (0.19)	(-0.53, 0.23)
	Indian	0.22 (0.24)	(-0.25, 0.69)
	Pakistani/Bangladeshi	0.88 (0.20)	(0.48, 1.28)
	Black	-0.48 (0.18)	(-0.84, -0.13)
	Other	0.42 (0.33)	(-0.23, 1.07)
Tenure (ref: own)	Rent	-0.068 (0.11)	(-0.28, 0.14)
	Other	-0.54 (0.18)	(-0.89, -0.18)
Accom. (ref: house)		-0.31 (0.12)	(-0.54, 0.080)
Educ. quals.	NVQ 2	-0.18 (0.15)	(-0.47, 0.11)
(ref: NVQ = 1)	NVQ 3	-0.024 (0.15)	(-0.32, 0.27)
	NVQ 4	0.18 (0.16)	(-0.14, 0.50)
	NVQ 5	0.35 (0.24)	(-0.12, 0.82)
	Overseas, none	-0.15 (0.15)	(-0.45, 0.15)
Child breast fed (ref: no)		0.28 (0.088)	(0.11, 0.46)
Main respondent in work (ref: no)		0.14 (0.075)	(-0.0094, 0.29)
Non-response to income qn. (ref: no)		-0.11 (0.12)	(-0.36, 0.13)
Wave non-response (ref: no)		-0.91 (0.090)	(-1.1, -0.73)
Participate in next sweep?	2	-0.58 (0.11)	(-0.79, -0.37)
(ref: 1 - very likely)	3	-0.94 (0.19)	(-1.3, -0.57)
	4	-2.0 (0.41)	(-2.8, -1.2)
	5	-1.4 (0.44)	(-2.3, -0.59)
Enjoyed IV?	2	-0.28 (0.09)	(-0.46, -0.10)
(ref: 1 – a great deal)	3	-0.50 (0.13)	(-0.76, -0.25)
	4	-0.82 (0.25)	(-1.3, -0.32)
	5	0.39 (0.41)	(-0.41, 1.2)

methods, data, analyses | Vol. 11(1), 2017, pp. 45-66 DOI: 10.12758/mda.2016.013

Collecting Event History Data with a Panel Survey: Combining an Electronic Event History Calendar and Dependent Interviewing

Josef Brüderl[1], *Laura Castiglioni*[1], *Volker Ludwig*[2], *Klaus Pforr*[3] & *Claudia Schmiedeberg*[1]

1 *LMU Munich*
2 *University of Kaiserslautern*
3 *GESIS*

Abstract

Many panel surveys collect event history data on events occurring between two waves. This is usually done by asking lists of questions on the various changes that took place between interviews (Q-Lists). Recently, some panel surveys introduced a different data collection method: the Event History Calendar (EHC), credited for collecting more accurate data. However, even the use of an EHC cannot prevent the issue that events tend to be reported spuriously at the seam of consecutive waves (seam effect). On the other hand, research has shown that dependent interviewing (DI) can help reduce this seam effect. Thus, the combination of EHC and DI (DI-EHC) promises to provide more accurate event history data that are not plagued by a seam effect. The German Family Panel pairfam was one of the first panel studies to use DI-EHC. In this article we first report on the practical aspects and the pros and cons of DI-EHC. Further, we report the results of an experiment in which we test whether DI-EHC reduces the seam effect. In sum our practical experiences and the results of our experiment indicate that the instrument is less burdensome than traditional Q-Lists and produces more accurate data. In particular, DI-EHC reduces the seam effect significantly.

Keywords: Event History Calendar; Dependent Interviewing; Seam Effect; Panel Survey; Questionnaire Design

1 Introduction

Panel surveys ask prospective questions about respondents' life situation and status at the time of the interview. When done repeatedly over several waves, this process produces panel data. In addition, many panel surveys also collect event history data by asking respondents retrospective questions regarding status changes such as transitions and events that occurred in the time since the last interview. Compared with classic panel data, such event history data allow for a more precise modelling of the timing of certain events (e.g., survival analysis). Traditionally, event history data have been collected by means of question lists (Q-Lists), looping over the statuses that have been reported by respondents and asking about the beginning and end time of each episode. These loops can move forward from the status at the last interview or backward from the current status.

However, retrospective reports of episodes can be biased by recall mistakes. A number of recall errors have been reported in the literature (Eisenhower, Mathiowetz, & Morganstein, 1991; Sudman & Bradburn, 1974). For instance, sometimes respondents do not report events or episodes altogether, leading to omission or underreporting of events. In other cases, timing errors such as telescoping or time expansion occur, i.e. reporting an event as having been more or less frequent than it actually was. These mistakes are a potential source of bias in event history data.

1.1 Event History Calendars

To improve the quality of retrospective data, calendar-based techniques – initially in form of paper-and-pencil calendars – have been suggested since the late 1960s as an alternative to Q-Lists (Balan, Browning, Jelin, & Litzler, 1969; Freedman, Thornton, Camburn, Alwin, & Young-DeMarco, 1988). Calendar instruments typically consist of a two-dimensional grid with the X-axis representing the timeline

Acknowledgments

This paper uses data from the German Family Panel pairfam, coordinated by Josef Brüderl, Karsten Hank, Johannes Huinink, Bernhard Nauck, Franz Neyer, and Sabine Walper. pairfam is funded as long-term project by the German Research Foundation (DFG). We are grateful to an anonymous reviewer, whose comments improved this paper very much. Further, we thank Madison Garret for language editing and proofreading.

Direct correspondence to

Josef Brüderl, Ludwig-Maximilians-Universität München, Institut für Soziologie,
Konradstr. 6, 80801 München, Germany
E-Mail: bruederl@lmu.de

(e.g. with months or years being the columns), and the Y-axis life domains such as employment or residences (with the respective statuses in place of the rows)[1]. Using this grid, respondents receive visual cues about the period on the timeline and can easily indicate for which cells of the grid an event or episode should be recorded. Landmark events such as birthdays or holidays can be included in the calendar to facilitate the timing of events. For retrospective surveys, calendar-based methods have become rather common (see the literature review provided in Glasner, 2011, p. 45). Since the late 1990s calendar instruments have also been introduced in electronic form in large panel surveys. In the panel context it has become common to term such instruments "Event History Calendars", or EHC (Belli, Stafford, & Alwin, 2009).

Many survey researchers argue that calendar instruments facilitate recall accuracy by means of a graphical presentation of timelines with visual cues that better fit respondents' idiosyncratic autobiographic memory structures (Belli, 1998). Furthermore, the conversational style of the interview improves respondents' recall (Belli, 2000; Caspi et al., 1996). Based on the graphical timeline, respondents are able to relate events to each other and detect gaps and inconsistencies in records (van der Vaart, 2004). For instance, landmark events can be used as temporal anchor points to which respondents can relate other events (e.g., "We moved to X the week before Christmas"). Similarly, multiple-domain calendars can help to link events across life domains (e.g., "We moved in together just before I graduated"). Accordingly, evaluations of calendar-based techniques have shown that calendar instruments improve data quality regarding completeness and consistency compared to data collection by means of question lists (Belli & Callegaro, 2009; Glasner & van der Vaart, 2009). Although the beneficial effects of calendars were found to be more important for recall of less recent events (Glasner & van der Vaart, 2009), they may be just as helpful for accurate reports of the relatively short periods between panel waves.

Data quality is increased further thanks to calendar instruments as they improve the interviewing process. The graphical representation of the information already recorded in the calendar renders detection of gaps and inconsistencies very easy for the interviewers, who receive cues to probe accordingly. For this reason, EHCs are also implemented in telephone surveys such as the PSID, where solely the interviewer, not the respondent, can see the calendar. A typical feature of calendar-based data collection is the greater degree of flexibility allowed to interviewers: they may deviate from the given question order and wording to help the respondent more accurately recall a series of episodes (Belli & Callegaro, 2009). Indeed, research has shown that interviewer variance is slightly increased by the use of EHC methods in a CATI survey (Sayles, Belli, & Serrano, 2010), which can be interpreted

[1] For a detailed description of characteristics of calendar-based data collection see Callegaro (2007).

as a sign of greater flexibility provided by this method. As a consequence, a conversational interaction is possible which may lead to higher motivation and reduce satisficing (Belli & Callegaro, 2009; Belli, Lee, Stafford, & Chou, 2004; Krosnick, Narayan, & Smith, 1996). Calendar instruments are in fact reported to be preferred by both respondents and interviewers over question lists (Freedman et al., 1988). A field test of the newly developed EHC in the re-engineered SIPP revealed that respondents perceived the calendar-based instrument as more interesting than the traditional interview (Chan, 2009). In the experimental comparison between EHC and Q-Lists conducted in the PSID 1998 Calendar Methods Study (Belli, Shay, & Stafford, 2001), interviewers reported to have enjoyed the EHC interviews more than traditional question lists.

1.2 Seam Effect

In the context of panel surveys, recall errors may produce a specific methodological problem: the so-called "seam effect". A seam effect means that we observe a higher rate of change at the seam between two consecutive panel waves than within the period a respondent reports on during the interview (Burkhead & Coder, 1985; for a review see Callegaro, 2008). Seam effects are the product of both the underreporting of transitions within a wave ("constant wave reporting") and spurious changes between waves (Jäckle, 2008; Rips, Conrad, & Fricker, 2003). In particular, a spurious change can occur if the respondent classifies the same status differently in two consecutive waves (misclassification). Another mechanism is omission: in this case, the last months of an ongoing episode from the previous calendar are "forgotten" in the next wave. Finally, due to backward telescoping transitions are often dated back to the seam. Thus, paradoxically, when collecting event history data via panel surveys we might minimize retrospective recall bias on the one hand; however, on the other we introduce artificially high transition rates at the seams.

Data collection using an EHC can help to decrease seam effects (Callegaro, 2007). Research has shown that calendar-based data collection methods are often superior to question lists with regard to underreporting or time error (Belli, Shay, & Stafford, 2001; Belli, Smith, Andreski, & Agrawal, 2007). Thus, as calendar instruments facilitate recall – e.g. due to the use of landmark events, visual cues regarding the temporal order of episodes, and the visibility of inconsistencies in entries – the accuracy of event history data will be improved and inconsistencies between waves will be less likely (Callegaro, 2007; Rips et al., 2003).

As a more specific method to tackle the seam problem, dependent interviewing (DI) has been introduced by a number of panel studies since the 1990s. Information from previous waves is preloaded to tailor the wording of questions (proactive DI), or for automatic consistency checks (reactive DI) (Callegaro, 2008; Jäckle, 2009). For instance, in proactive DI, instead of recording the employment status at the beginning of the reference period, the interviewer asks if the respondent has maintained the same employment status recorded in the previous interview ("according to my records, last year you …, is this still the case?"). In reactive DI, automatic consistency checks may highlight if the status reported for the same point in time in the previous wave differs from that reported at the current interview. In this case, interviewer and respondent can revise the data together to solve the inconsistency.

DI has been proven an effective method to reduce seam effects (Jäckle & Lynn, 2007; Moore, Bates, Pascale, & Okon, 2009) as preloads reduce both the chance of misclassification and omission (Lynn, Jäckle, Jenkins, & Sala, 2012; Lynn & Sala, 2006). Further, the problem of backward telescoping should be minimized: transitions cannot be dated back to the seam as the preloaded status must first come to an end.

1.3 DI-EHC

Building upon this knowledge, it seems promising to combine EHC and DI (DI-EHC) as a means to increase recall accuracy and reduce seam effects (Callegaro, 2008). While DI reduces spurious change between waves due to misclassification, omission, or backward telescoping, EHC may help reduce constant wave reporting and underreporting of short or seemingly irrelevant episodes. One of the aspects of DI which could be potentially problematic is that DI can trigger cognitive satisficing (see Krosnick, 1991): respondents might feel that their interview is easier and shorter if they confirm the data prompted by the preload. However, by implementing DI in a calendar setting the pairfam questionnaire does not offer any strong incentives for confirming the preload throughout the reference period as the interviewers need the same amount of time whether they check off one category or another[2]. Hence, also in this respect, the combination of EHC and DI might trigger positive synergies between the two methods.

The German Family Panel pairfam has introduced a DI-EHC for collecting data on partnerships, residences, education, and employment. The aim was to improve recall accuracy and to reduce the seam effect. Due to the lack of validation data, we cannot investigate whether the accuracy of the data increased. However, we can investigate whether the seam effect increased among a randomly chosen subgroup of the respondents for whom we experimentally excluded preload data in the education and employment calendar compared to the majority of respondents for whom all preloads were included. Therefore, the main purpose of this article is

[2] Hoogendoorn (2004) found that the issue of acquiescence in connection with proactive DI can be solved by certifying that confirming the preloads would not translate to a sizable shortening of the questionnaire. Also Eggs and Jäckle (2015) and Jäckle and Eckman (2016) have found no support for the hypothesis that proactive DI leads respondents to satisfice.

to report the results of our randomized methods experiment on the effectiveness of DI-EHC for reducing the overall seam effect.

1.4 Contents of the Paper

This paper is structured as follows: First, we will give an overview of EHC modules in existing panel studies. Then, we will describe the structure of the DI-EHC in pairfam and practical aspects of its implementation. As Glasner and van der Vaart (2007) point out, in recent years calendar instruments were developed without taking advantage of experiences made in other studies. With our overview and the practical guide to the pairfam DI-EHC we hope that other studies might learn from our own experiences. The results of our experiment will follow. Finally, we conclude and discuss lessons learned and give recommendations for future developments.

2 EHC in Other Panel Studies

The Survey of Income and Program Participation (SIPP) was among the first panel studies to use calendar techniques. In fact, the first considerations when introducing calendar-based data collection in the SIPP already aimed at eliminating the seam effect and included DI techniques (Kominski, 1990). Interviewers filled out a graphical paper-and-pencil calendar after the first interview and handed it over to the respondent at the beginning of the second interview. After the second wave interview, the interviewer updated the calendar and gave it to the respondent again at the beginning of the third interview. Although the interviews were conducted using conventional question lists, respondents could use information displayed in the calendar from the previous waves as well as visual cues when answering the retrospective questions. This early EHC was implemented to aid respondents rather than as a data collection instrument itself: data were still collected by standard question lists, and the paper-based calendar distributed to respondents was used only to illustrate data entries from previous waves as a mere recall aid. A further step was taken in 2007 when a computer-assisted EHC was designed as an integral part of the survey. The reason for this development was the decision to change from the former design of three interviews per year to an annual survey (Fields & Callegaro, 2007). This shift raised concerns about respondents' ability to accurately report over this longer period. After field tests in 2008 (Chan, 2009; Pascale, 2009) and 2010 (Moore, 2012), the re-engineered SIPP including the computerized EHC was finally fielded in the 2014 SIPP Panel.

The Panel Study of Income Dynamics (PSID) also implemented an EHC when the interview cycle was changed from annual to biennial interviews (Beaulé, Dascola, & Liu, 2009). The “1998 PSID Calendar Methods Study” was conducted to compare the quality of data collected using the EHC versus standardized question list methods (Belli et al., 2001), but it was not until 2003 that the PSID employment module was reprogrammed as an EHC (Belli et al., 2007). As the PSID is a telephone survey, the EHC was only designed to help the interviewer detect inconsistencies such as gaps in employment history and overlaps in employment spells (Beaulé et al., 2009). The calendar spans a 2-year period and is rather detailed, with a third-of-a-month as the smallest unit (Belli et al., 2007). It contains the five following life domains: landmark events, residence, employment, not working, and time away. All domains were visible on one screen with separate summary timelines for each to facilitate parallel retrieval. Programmed consistency checks helped the interviewer detect potential inconsistencies. The experiences of the PSID team showed that by using the calendar method, post-processing time could be reduced (Beaulé et al., 2009).

A similar approach was taken by the adult cohort of the National Education Panel Survey (NEPS) in Germany. As it is also a telephone survey, event history data on education and employment are collected via Q-Lists using DI. This Q-List module is followed by a calendar-based data-revision module: the survey software automatically reorganizes all entries into calendar form in order to support the interviewer in correcting inconsistencies and in detecting biography gaps (Drasch, Kleinert, Matthes, & Ruland, 2016; Trahms, Matthes, & Ruland, 2016).

Other studies use simple calendars for single domains, for instance the Household, Income and Labour Dynamics in Australia Survey (HILDA, (Watson, 2009)) and the German Socio-Economic Panel (SOEP). We will not describe these calendars in detail here.

To summarize, several large scale panel studies have been combining DI and Q-Lists. Several studies have also used EHCs, albeit mainly for the purpose of data editing. However, to our knowledge thus far only one large scale panel study –pairfam – has implemented a combination of DI and EHC.

3 DI-EHC in the German Family Panel (pairfam)

The German Family Panel pairfam (Huinink et al., 2011) is a multi-disciplinary, longitudinal study on partnership and family dynamics in Germany based on a nationwide random sample of initially more than 12,000 persons of the three birth cohorts: 1971-73, 1981-83, 1991-93. Starting in 2008 the panel study collects data in annual waves via computer-assisted personal interviews (CAPI) administered by professional interviewers.

The purpose of the study is to collect comprehensive data on respondents' intimate relationships and family life, as well as social and economic circumstances. Research topics include partnership formation, institutionalization of intimate relationships, family formation and parenthood, and separation and divorce. For such research questions, accurate data regarding the temporal ordering of events including the start of a relationship, moving in together, marriage, or separation are crucial. Therefore, one of the core features in pairfam is an EHC on intimate relationships, places of residence, and occupations (i.e., school enrollment and labor force participation) spanning the period between the previous and current interview.

The EHC in pairfam has several unique features (presented in more detail below) to ensure high quality data. When developing the calendar we considered both theoretical findings on memory structure (Belli, 1998) as well as existing instruments (e.g. Belli et al., 2007; Reimer & Matthes, 2007), adapting them to the specific interests and needs of the pairfam study. In particular, the pairfam EHC incorporates DI techniques and implements an individually adjusted calendar span: the starting month is that of the last interview, and the maximal duration is set to 32 months to accommodate for one wave non-participants.

The EHC is a stand-alone Java application that is fully integrated into the pairfam interview. This was done for the higher flexibility regarding graphic interfaces offered by Java compared to the available CAPI software[3]. The EHC starts as a pop-up window after some "warm-up" questions. It displays some information we feed forward from the previous wave(s), the so-called "preloads". At the end of the EHC module the collected data are stored in the main dataset of the CAPI-software so that all entries are available in the following part of the interview for routing[4].

Interviewers allow the respondent to observe the screen while completing the EHC to ensure that both the interviewer and the respondent profit from the graphic representation of the entries. For each calendar a scripted introductory text is given while further probing in the case of gaps and inconsistencies is not scripted. In addition, questions are scripted for each line in the calendar ("In which months were you together with [name of the partner]?").

For illustration purposes, we present here the introduction question in the case of the partnership calendar:

> *"We would now like to know in which months you and your partner were in a relationship, if and when you were living together, or were married. You can see here a calendar with one column for each month between the last interview and today. Also, your current partnership status is entered in the column*

[3] The pairfam questionnaire was initially programmed in IN2Quest until wave 3. From wave 4 onwards, the questionnaire runs on a NIPO CAPI software. Both programs offer limited support for tailoring the graphic of the questionnaire interface.

[4] The EHC is designed to be fully functional also without preloads. This is to facilitate the implementation of a refreshment sample.

> *labelled 'now'. For each partner there are three lines, one for having a relationship, one for living together, and one for marriage or civil union. We will proceed as follows: You will look at the screen and tell me what has happened since the last interview [*first-time respondents: *"EHC time period"], and I will enter the data. After we are finished, you can check if the information I have entered is correct".*

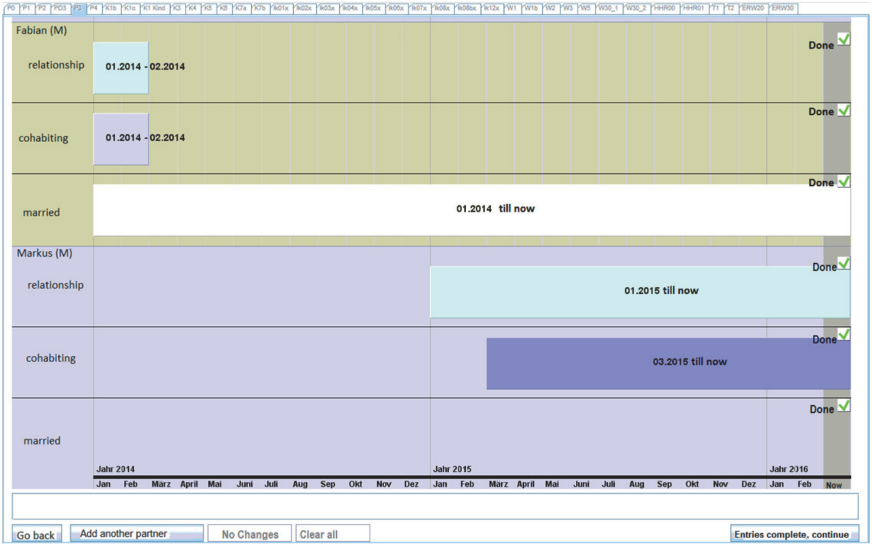
The EHC covers three life domains: intimate relationships (including cohabitation and marriage), places of residence, and education and/or employment episodes. Furthermore, two "synoptic calendars" conclude the EHC in order to crosscheck entries before the EHC is closed and the normal interview continues. A detailed description of the EHC including specific wording of questions and consistency checks in pop-up windows can be found in the pairfam codebooks (e.g. pairfam, 2015).

All calendars enable monthly entries and cover the time span since the last interview. Therefore, as the period between two interviews varies between respondents the length of each calendar is adjusted for each respondent individually. The last available information (column "now") from the previous interview is used to produce the preloads for the next wave.

In addition to the monthly entries the pairfam EHC includes one further column for the current situation (column "now") in order to take into account the most recent changes in respondents' lives. This is particularly important as information from the EHC is subsequently used for routing purposes in the remainder of the interview. For instance, let us assume that the interview takes place in January and the respondent reports that their relationship ended in January, too. Some respondents might then say that January is the last month in a relationship (for instance, if they split up towards the end of the month) and hence should be marked as such in the calendar. However, for the following section of the interview we will want to save the information that the respondent is currently single. For this reason, the interviewer will enter January as the last month in a relationship and make sure that the cell corresponding to 'now' of the respective row is unmarked.

Once all data are entered, a box at the end of each row in the calendar must be ticked in order to show that the row has been completed. The check mark disappears again if the interviewers alter any entry in this row. This feature was implemented to ensure that interviewers notice unintentional changes of the record.

The first life domain covered in the EHC is intimate relationships. Other calendars (e.g. SIPP) often begin with respondents' residential mobility, since moving is a rather seldom event and moving dates tend to be easy to recall. In the pairfam study we decided to start with a life domain that is more central to our study as we thought that this might help to keep the attention high. The first screen requires that respondents list the names of the partners with whom they were together since the last interview (partner list). If a respondent had a partner in the last interview his/



Fabian (M)

relationship: 01.2014 - 02.2014 Done ✓

cohabiting: 01.2014 - 02.2014 Done ✓

married: 01.2014 till now Done ✓

Markus (M)

relationship: 01.2015 till now Done ✓

cohabiting: 03.2015 till now Done ✓

married: Done ✓

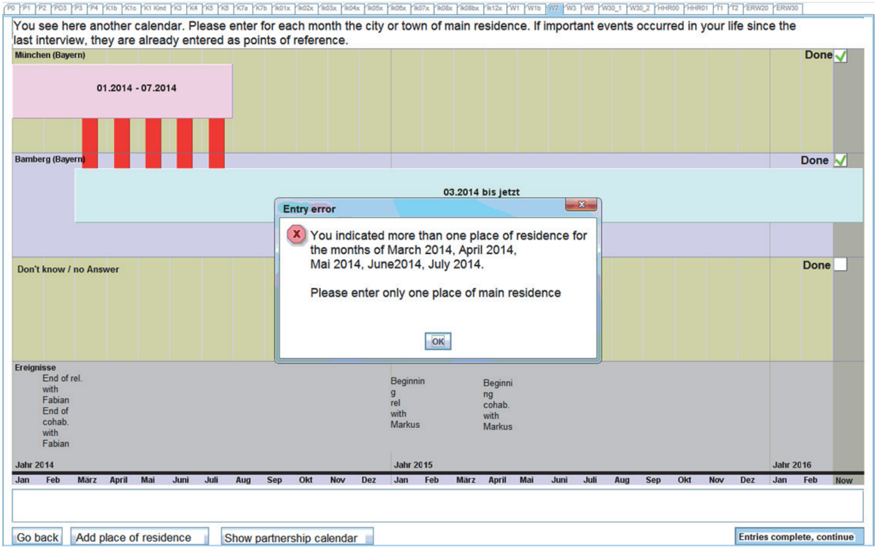
Figure 1 Partnership calendar containing two relationships: the respondent was married and living together with her partner at the time of the previous interview, is now still legally married but lives together with a new partner.

her name, gender, and date of birth is already included in that list as a preload. This introductory screen was not introduced for a methodological but rather for a technical reason: knowing the number of partners prior to opening the calendar helps optimizing the height of the rows in the calendar, as the calendar always contains only as many rows as necessary for the number of partners mentioned.

After the partner list is completed, the partnership calendar is shown (Figure 1). All names entered in the list appear automatically in the calendar view. First the partner from the last interview is listed, then, where necessary, new partners. For each of the partners there are three rows: the first one reports in which months the relationship existed, the second one in which months the respondent and their partner lived together, and the last one is for reporting marriage duration. For the preloaded partner the cells of the interview month of the last interview are marked according to the information given in the column "now" in the last interview. Thus, we preload information on the status of the partnership also.

To avoid incorrect entries, the calendar includes a number of consistency checks which are run as soon as the interviewer declares the data entry to be complete for this life domain. For instance, if parallel marriage episodes (i.e. with two partners) are entered, a pop up window will indicate an inconsistency that requires a correction. In other cases such as parallel cohabitation episodes, the consistency check only triggers a pop-up window but correction is not required, as this is a rare but possible arrangement. Additionally, if respondents indicate that a relationship has ended, a pop-up window appears with a question as to whether the relationship ended due to separation or death of the partner. Similarly, in case of a new marriage a pop-up window asks if the wedding ceremony was religious or civil.

Like the partnership calendar, the following residence calendar is preceded by one question recording all the places of residence in which the respondent has lived since the past interview (residence list). For each place of residence the municipality and the federal state are entered. The place of residence at the time of the last interview is already included at the top of the list as a preload variable. The design of the residence calendar is similar to the partnership calendar: each place of residence listed in the introduction question is assigned to a row of the calendar table. The preloaded place of residence is displayed in the upper row and the month of the past interview is already marked (preload). For each place of residence the interviewer marks the months in which the respondent lived there (see Figure 2). Gaps are not acceptable: the respondents must indicate a place of residence for each month. Overlaps of two places of residence of not more than one month are allowed in order to account for moves within a month. If respondents wish to add a further place of residence during calendar completion, a button adds a further row without turning back to the residence list. In addition, as respondents might have difficul-



You see here another calendar. Please enter for each month the city or town of main residence. If important events occurred in your life since the last interview, they are already entered as points of reference.

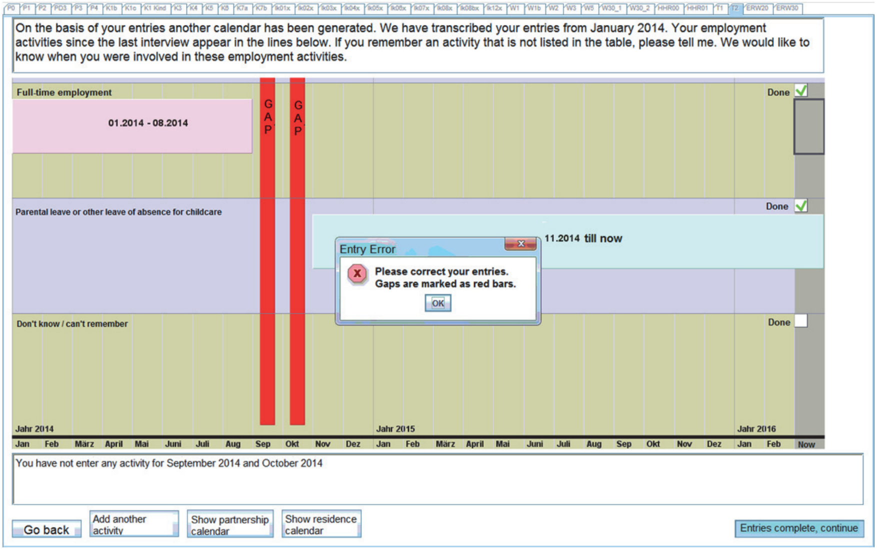
Entry error

You indicated more than one place of residence for the months of March 2014, April 2014, Mai 2014, June2014, July 2014.

Please enter only one place of main residence

P0 P1 P2 PD3 P3 P4 K1b K1c K1 Kind K3 K4 K5 K6 K7a K7b Ik01x Ik02x Ik03x Ik04x Ik05x Ik06x Ik07x Ik08x Ik08bx Ik12x W1 W1b W2 W3 W5 W30_1 W30_2 HHR00 HHR01 T1 T2 ERW20 ERW30
München (Bayern)
01.2014 - 07.2014
Done ✓
Bamberg (Bayern)
03.2014 bis jetzt
Done ✓
Don't know / no Answer
Done □
OK
Ereignisse
End of rel. with Fabian
End of cohab. with Fabian
Beginnin g rel with Markus
Beginni ng cohab. with Markus
Jahr 2014
Jahr 2015
Jahr 2016
Jan Feb März April Mai Juni Juli Aug Sep Okt Nov Dez Jan Feb März April Mai Juni Juli Aug Sep Okt Nov Dez Jan Feb Now
Go back
Add place of residence
Show partnership calendar
Entries complete, continue

Figure 2 Residence calendar. The respondent entered parallel residence spells for five months instead of marking only the main place of residence



On the basis of your entries another calendar has been generated. We have transcribed your entries from January 2014. Your employment activities since the last interview appear in the lines below. If you remember an activity that is not listed in the table, please tell me. We would like to know when you were involved in these employment activities.

Full-time employment — Done ✓

01.2014 - 08.2014

Parental leave or other leave of absence for childcare — Done ✓

11.2014 till now

Don't know / can't remember — Done □

Entry Error

Please correct your entries.
Gaps are marked as red bars.

OK

You have not enter any activity for September 2014 and October 2014

Figure 3 Activity calendar. The respondent entered no activity for two months, which are marked red.

ties to align the timing of a move with major events indicated in the partnership calendar, beginning and end time of relationships and children's dates of birth are displayed at the bottom of the calendar.

Finally, the calendar for education and employment (activity calendar) is preceded by a list of 22 possible activities (activity list). In contrast to the first two calendars, the activity list is not prefilled with preloads. The activities reported in the previous interview are displayed afterwards in the calendar together with the activities ticked in the activity list. Thus, concerning the activity calendar DI is not used in the first step, when collecting last years activities, but only in the second step, when filling out the calendar. It turned out that this design decision was suboptimal (see below).

The activity calendar (Figure 3) contains a row for each of the relevant activities. If no activity is ticked in the activity list, the row "don't know" appears in the calendar (in addition to the rows of the preloaded activities). Also in this case some crosschecks are programmed to ensure that unlikely combinations of activities cannot be entered by mistake. Gaps are not acceptable: months with no status are marked red in the calendar and a pop-up window lists all months with no information.

Our experiences with the EHC are positive. Most parts of our DI-EHC work smoothly and deliver plausible data. The residence calendar was the only one which required a few structural changes in the first years of its implementation as we decided to reduce the level of precision of our residential history by focusing only on the primary place of residence. In the first years, we required the respondents to enter also their secondary residence places, but this effort did not pay off: the data collected were often contradictive and many changes turned out to be spurious. After giving up a comprehensive data collection of secondary residences, we opted for a simple question as to whether respondents have a second residence. This information is necessary for subsequent questions regarding respondents' mobility, for example. Further minor adjustments were necessary in the partnership and children modules. We introduced additional checks to avoid preloads being deleted by mistake and now require the interviewer to enter a reason for deleting partners or children.

Interviewers are used to standardized question lists but not to the more flexible calendar-based instruments. Therefore, interviewers were made acquainted with the EHC prior to the field start of wave 2 – the first one with an EHC. Nonetheless, the pairfam team discovered that a certain number of typical coding errors had occurred in the first waves after implementation. Preload deletions occurred particularly often. In subsequent waves some of these errors could be eliminated by implementing additional pop-ups. To further improve data quality we also introduced an interview rehearsal in wave 4. Before the start of each wave, a fictive case is constructed with a large number of (more or less complicated) events and transitions during the period covered by the EHC. Interviewers receive a written description of the case and have to record this fictional interview in the EHC. From the data produced the project team can examine the errors made and specifically address those issue both during interviewer training and in the interviewer handbook. Interviewers who made too many errors receive additional training. After the introduction of this rehearsal interview the number of coding errors decreased. In the following waves, we tailored the description of the fictive case to address specific concerns and recurrent mistakes detected during data cleaning.

Producing preloads for the next wave is quite demanding: Each year at the end of May the survey agency delivers the raw data of the last wave and preloads are needed by the end of September before the fieldwork of the next wave begins in October. We feed forward more than 300 preload variables which need to be validated for plausibility and, in the case of data such as names, places, and job descriptions, must be checked for spelling errors. Preloads must be prepared carefully as mistakes can cause unpleasant incidents during the interview.

4 Does DI-EHC Reduce the Seam Effect? Results from a Randomized Experiment

In order to investigate the effectiveness of the pairfam DI-EHC in reducing the seam effect, we implemented an experiment in wave 3 of the survey: we randomly selected 1,000 (11%) of the 9,069 wave 2 respondents and deleted their preloads before fieldwork started. We decided to limit the experiment to the activity calendar (education and employment status). In wave 3, 7,383 respondents from wave 2 could be re-interviewed. For 813 of these respondents (11%) no preloads where shown in their activity calendar. The other 6,570 respondents got a complete set of preloads[5].

Calendar data collected in wave 3 were matched with those from wave 2 in order to analyze transitions in educational or employment status at the seam from wave 2 to wave 3. We use the monthly information on respondents' status (variables ehc19i$m1-ehc19i$m18). We set up a long format panel data set where each row is a person-month. The data cover all months from the wave 1 interview to the wave 3 interview[6]. For the wave 2 interview month (MonthIntW2) three pieces of information on the activity status are available: status in MonthIntW2 as collected in wave 2, current status also collected in wave 2 (ehc19i$), and status in MonthIntW2 as collected in wave 3 (ehc19i$m1). For transition analyses on a monthly basis one has to decide, which status information should represent MonthIntW2. We decided to use the information on the current activity status that is recorded in the wave 2 interview (ehc19i$). Individual panels are organized such that the seam between waves 2 and 3 is at month 0 and up to 17 preceding and 17 following months are available. Note that due to varying wave distance, the number of person-months varies across persons.

Our main outcome in the experiment is the proportion of respondents reporting any change in status between two ensuing months t and t-1. We expect that respondents report more changes at the seam (that is, between the month of the wave 2 interview and the following month) than off-seam (any other month). However, the seam effect should be smaller for respondents who do see preloaded calendar information during the interview compared to respondents whose preloads were deleted as part of the experiment.

Our analysis is based on pairfam data release 6.0 (Brüderl, Hank et al., 2015). More details on the study can be found in Huininik et al. (2011). We decided to use

[5] Due to the experimental design, treatment assignment took place before fieldwork started. Hence, the cases interviewed were less than those originally selected. However, attrition rates are very similar across experimental groups (19.6% drop-out with preloads and 19.7% without).

[6] Employment status at the wave 1 interview month is not included in the wave 2 data. Due to a programming error, the information has not been recorded (variables ehc19i$m1 are empty in wave 2).

edited and released data (not exactly the raw data) as we are interested in the seam effect in the data actually available for research. However, the pairfam data team applied only minor changes to education and employment histories (Brüderl, Hajek et al., 2015).

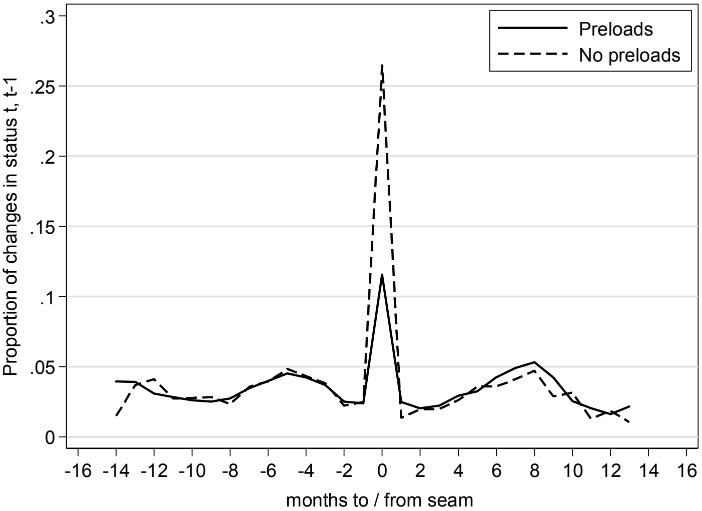
In the analyzed data there are 6,569 respondents with preloads (treatment group) and 813 respondents without preloads (control group). The two experimental groups provide 173,807 and 21,513 person-months, respectively. Note that one person belonging to the control group had to be excluded because the event history was invalid (as identified during the data cleaning process). Further, we excluded all person-months without any information on status (gaps in histories). Together, these restrictions eliminated 7,272 person-months (4.2%) from the treatment group and 912 person-months (4.2%) from the control group.

As the outcome of our analysis is defined as status changes between two months t and t-1, the final number of cases in the data set is smaller. For the earliest month of each respondent, a change is not defined. Furthermore, change is not defined for gaps in individual panels. The analysis of the proportion of changes in status is therefore limited to 159,831 (treatment group) and 19,773 person-months (control group). On average, repondents provide information on status changes for 24.3 months in both groups.

Results of our analysis are presented in graphical form in Figure 4. There is clear evidence of a seam effect in both groups. As expected, however, the seam effect is much smaller with preloads. Thus, our experiment demonstrates that using preloads substantially and significantly (see below) reduced the seam effect in the monthly education and employment histories in pairfam. Further, there seem to be no systematic differences between treatment groups off the seam. Thus, as intended, the treatment (preload information) reduces only artificial seam changes, but not "real" changes off the seam.

For taking a closer look, Table 1 shows sample proportions of monthly status changes on and off the seam for the control group (columns (1) and (2)) and for the treatment group (columns (4) and (5)). The difference in the proportion of changes on and off the seam gives us estimates of the seam effect without preloads (column (3)) and with preloads (column (6)). The difference in the seam effects between control group and treatment group then is our estimate of the treatment effect (column (7)). It tells us to what extent using preloads reduces the seam effect present in the EHC data. For the pooled sample of all three pairfam birth cohorts (first row of Table 1), the seam effect is 23.4 percentage points without preloads, but only 8.3 percentage points with preloads. Hence, providing respondents with information preloaded in the EHC substantially and significantly reduced the seam effect by 15.1 percentage points.

Table 1 also shows separate analyses for the three cohorts (born 1991-93, 1981-83 and 1971-73). For respondents from the youngest cohort, who in many cases



Note: Proportion of respondents reporting any change in activity status between months t and t-1. In each month, status can be any of 22 categories, including multiple statuses (9 categories for education, 7 categories for employment, 6 categories for non-employment; see pairfam anchor Codebook (pairfam 2015)). The figure is restricted to a maximum of 14 months before the seam and a maximum of 13 months after because due to the annual waves the number of observations outside this interval is low.

Source: pairfam release 6.0, anchor data waves 2 and 3.

Figure 4 Proportion of cases reporting a change of activity status compared to the previous month. Time line centered at the month following the wave 2 interview.

completed secondary schooling during the time observed here, the seam effect was smaller than for the two older cohorts (see Table 1). However, all results point in the same direction: for each cohort, we found a strong seam effect, which was significantly higher if preloads were deleted in the experiment.

As Figure 4 suggests, there are hardly any systematic differences in the proportions of status changes off the seam. For the pooled sample, average proportions off the seam are .031 without preloads and .033 with preloads. This difference is not significant at reasonable levels (p=.17). (There are also no significant differences when looking at birth cohorts separately.) Obviously, random assignment to the experimental groups worked well and, as desired, preloads reduced reported status changes only on the seam, but not off the seam. We further compared proportions between treatment and control group for each single month before and after the seam. We found a significant difference only in one instance, namely for the month following the seam. In this case, the proportion of changes was slightly larger for respondents with preloads than for those without preloads (.0247 com-

Table 1 Proportion of respondents reporting a change of status compared to the previous month by treatment group and estimated seam effects with and without using preloads

	No preloads			Preloads			
	Off seam	On seam	Seam effect	Off seam	On seam	Seam effect	Treatment effect
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
All 3 cohorts	.031	.265	.234*** (14.98)	.033	.116	.083*** (20.85)	-.151*** (-9.35)
Cohort 1991-93	.032	.205	.173*** (7.59)	.035	.096	.061*** (10.55)	-.112*** (-4.77)
Cohort 1981-83	.040	.308	.269*** (8.73)	.040	.135	.094*** (11.71)	-.174*** (-5.41)
Cohort 1971-73	.021	.299	.277*** (9.76)	.023	.125	.102*** (13.93)	-.175*** (-5.96)
Persons	813	805	813	6,569	6,529	6,569	7,382
Person-months	18,968	805	19,773	153,302	6,529	159,831	179,604

Notes: Seam effects are calculated as the difference in proportions On seam – Off seam; No preloads: (3) = (2) - (1); Preloads: (6) = (5) - (4). Treatment effect is the difference-in-differences estimator; (7) = (6) – (3). Two-sided tests for significant differences in proportions, adjusted for clustering of persons; z-values in parentheses. *** p<.001.

Source: pairfam release 6.0, anchor data waves 2 and 3.

pared to .0136). The difference of 1.1 percentage points is significant at the 5 percent level (z= 1.96; p=.05). This finding is consistent with backward telescoping where respondents without a preload date the event back to the first month (seam) of the calendar.

Given the results from the experiment, the question arises, why there still is a seam effect despite using preloads? In the pairfam case, the reason might be that preloads were fed into the calendar only, but not into the introductory question list (as we reported in the descpription of the employment and education calendar, see Section 3). So, when the calendar was first shown to the respondents, they had already given information without having seen the preloads. This design feature is suboptimal, because without preloads respondents might misclassify their activity status in the month of the last interview (a common example is "part-time employment" classified as "marginal employment", or the other way around). Respondents then see in the calendar the status that they reported in the last interview (preload) and the (misclassified) status that they reported just before in the activity list. Surely not all respondents will then delete the misclassified status and continue with the preloaded status (then we would observe no spurious change at the seam). Instead, quite a few respondents probably ignore the preload and continue the calendar with the misclassified status. Such respondent behavior will produce an artificial change on the seam. A close inspection of the data produced by the DI-EHC showed that this “misclassification mechanism” is indeed a source of the seam effect in the pairfam activity calendar (results not shown, but available upon request).

5 Discussion and Conclusion

In sum, the results of the experiment show that pairfam successfully reduced the seam effect by using DI-EHC. Nevertheless, a sizable seam effect still remains even with DI. These findings are consistent with earlier research by Jäckle and Lynn (2007) which showed that proactive DI substantially reduced seam effects in monthly work histories, but did not eliminate them completely. In the pairfam case, most likely a seam effect remains due to the fact that preloads were not used by design when collecting last year’s activities on a first screen.

This gives some hints how future research could improve on our results. Basically the design used in the pairfam activity calendar is only “partial DI-EHC”. By showing preloads only on the second screen (the activity calendar) two mechanism producing the seam effect could be alleviated: omission and backward telescoping. However, the third mechanism – misclassification – still operated, because preloads were not used on the first screen, when a list of activities was shown. Therefore, we speculate that most of the remaining seam effect is due to misclassification. This could be investigated by designing an improved experiment with a third experimental group added that gets a “full DI-EHC” (preloads used on both the activity list and the activity calendar).

In addition, there are some more practical aspects when using DI-EHC. First, using preloads might also reduce interview duration. In pairfam the duration of each section of the questionnaire was recorded. In particular, the duration of each section of the EHC was tracked: The mean duration of the activity calendar was 1.36 minutes with preloads and 1.49 minutes without preloads. The difference is statistically significant ($p<0.01$). This result is particularly welcome as previous literature reports calendar interviews as such to be longer than Q-lists (Glasner & van der Vaart, 2009, p. 63).

Second, feedback from interviewers suggests that this instrument is less tiresome than question lists both for interviewers and respondents. In particular this is, because DI-EHC avoids boring repetitions.

Finally, there are a few things to bear in mind in order to achieve high-quality results with this method. Firstly, the effort necessary to have an effective and appealing instrument should not be underestimated. The pairfam team outsourced the programming to the field agency, and the actual software development started almost one year before the beginning of fieldwork. Even so, this proved to be a tight schedule. Our objectives in terms of flexibility and appeal would have required more specific programming and human interface design skills. Some of our aims (e.g., the parallel visualization of all three calendars) could not be achieved with the resources at our disposal. Remodeling the calendar is also quite resource intensive and in panels not desirable in order to ensure comparability across waves. Hence, it is mandatory to invest enough time in the conception phase.

One big advantage of an EHC is the possibility to implement rather complex consistency checks during data collection. Possible mistakes can be defined quite easily upfront by survey managers (e.g., pop-ups) and can be immediately communicated and corrected during the interview. In our experience, adding additional pop-ups to cross check improbable entries is rather simple, and facilitates avoiding accidental data entries/deletions.

Before fielding a survey with an EHC module, interviewers must be trained extensively to properly use this instrument. Using a partially scripted questionnaire is challenging, especially if they have never done it before (for a vivid illustration on what interviewers (and respondents) mess up when using such complex instruments see Uhrig and Sala, 2011). Furthermore, we found that not all interviewers were comfortable with the graphic interface. It is advisable to gain a good grasp of common mistakes and make sure that interviewers learn how to properly navigate the calendar (rehearsal interviews are a very effective method).

Finally, an EHC produces a large amount of information. For each life domain the status for every month is recorded. These are sequence data on the interval since the last interview. These "pieces" of the life-course must be consolidated in some kind of biographical data set. Often this is also done in an episode format to facilitate event-history analyses. This process is very demanding and requires a lot of manpower, especially in the first couple of panel years when the data cleaning procedures are still in development.

All in all, the setup costs of a DI-EHC are not negligible: development, programming, interviewer training, and data handling procedures will require more resources than with a traditional CAPI. Nevertheless, in the long run, costs reduce to the level of a standard interview. On the other side, data quality is improved from the beginning and is even likely to improve with each further wave. Hence, the longer the planned duration of a longitudinal study, the higher the rate of return from a DI-EHC.

References

Balan, J., Browning, H. L., Jelin, E., & Litzler, L. (1969). A computerized approach to the processing and analysis of life histories obtained in sample surveys. *Behavioral Science, 14*(2), 105-120.

Beaulé, A., Dascola, M., & Liu, Y. (2009). *Development and Programming of an Employment Event History Calendar in the Panel Study of Income Dynamics* (PSID Technical Paper Series No. 09-01).

Belli, R. F., Smith, L. M., Andreski, P. M., & Agrawal, S. (2007). Methodological Comparisons Between CATI Event History Calendar and Standardized Conventional Questionnaire Instruments. *Public Opinion Quarterly, 71*(4), 603-622. doi:10.1093/poq/nfm045

Belli, R. F. (1998). The Structure of Autobiographical Memory and the Event History Calendar: Potential Improvements in the Quality of Retrospective Reports in Surveys. *Memory, 6*(4), 383-406. doi:10.1080/741942610

Belli, R. F. (2000). Computerized Event History Calendar Methods: Facilitating Autobiographical Recall. *Proceedings of the Survey Research Methods Section, ASA,* 471-475.

Belli, R. F., & Callegaro, M. (2009). The emergence of calendar interviewing: A theoretical and empirical rationale. In R. F. Belli, F. P. Stafford, & D. F. Alwin (Eds.), *Calendar and time diary methods in life course research* (pp. 31-52). Thousand Oaks: SAGE Publications Inc.

Belli, R. F., Lee, E. H., Stafford, F. P., & Chou, C.-H. (2004). Calendar and question-list survey methods: Association between interviewer behaviors and data quality. *Journal of Official Statistics, 20*(2), 185.

Belli, R. F., Shay, W. L., & Stafford, F. P. (2001). Event history calendars and question list surveys: A direct comparison of interviewing methods. *Public Opinion Quarterly, 65*(1), 45-74.

Belli, R. F., Stafford, F. P., & Alwin, D. F. (Eds.). (2009). *Calendar and time diary methods in life course research*. Thousand Oaks: SAGE Publications Inc.

Brüderl, J., Hank, K., Huinink, J., Nauck, B., Neyer, F. J., Walper, S., . . . Wilhelm, B. (2015). The German family panel (pairfam): ZA5678 Data file Version 6.0.0. *GESIS Data Archive, Cologne*. doi:10.4232/pairfam.5678.6.0.0.

Brüderl, J., Hajek, K., Herzig, M., Huyer-May, B., Lenke, R., Müller, B., . . . Schumann, N. (2015). *pairfam Data Manual: Release 6.0*. Munich.

Burkhead, D., & Coder, J. (1985). *Gross Changes in Income Recipiency from the Survey of Income and Program Participation* (Proceedings of the American Statistical Association, Social Statistics Section No. 351-356). Washington, D.C.

Callegaro, M. (2007). *Seam Effects Changes Due to Modifications in Question Wording and Data Collection Strategies: A Comparison of Conventional Questionnaire and Event History Calendar Seam Effects in the PSID*. Lincoln, NE: University of Nebraska.

Callegaro, M. (2008). Seam Effects in Longitudinal Surveys. *Journal of Official Statistics, 24*(3), 387-409.

Caspi, A., Moffitt, T. E., Thornton, A., Freedman, D., Amell, J. W., Harrington, H., . . . Silva, P. A. (1996). The life history calendar: a research and clinical assessment method for collecting retrospective event-history data. *International Journal of Methods in Psychiatric Research, 6*(2), 101-114.

Chan, A. Y. (2009). *The 2008 SIPP Event History Calendar (EHC) Field Test: Respondents' Reactions to the Interview* (Statistical Research Division Research Report Series (Survey Methodology) No. 03). Washington, D.C.

Drasch, K., Kleinert, C., Matthes, B., & Ruland, M. (2016). Why Do We Collect Data on Educational Histories Over the Life Course the Way We Do ? Core Questionnaire Design Decisions in Starting Cohort 6 – Adults. In H.-P. Blossfeld, J. von Maurice, M. Bayer, & J. Skopek (Eds.), *Methodological Issues of Longitudinal Surveys: The Example of the National Educational Panel Study* (pp. 331–347). Wiesbaden: Springer Fachmedien Wiesbaden. Retrieved from http://dx.doi.org/10.1007/978-3-658-11994-2_19

Eggs, J., & Jäckle, A. (2015). Dependent Interviewing and Sub-Optimal Responding. *Survey Research Methods, 9*(1), 15. doi:10.18148/srm/2015.v9i1.5860

Eisenhower, D., Mathiowetz, N. A., & Morganstein, D. (1991). Recall error: Sources and bias reduction techniques. In P. P. Biemer, R. M. Groves, L. E. Lyberg, N. A. Mathiowetz, & S. Sudman (Eds.), *Measurement Errors in Surveys* (pp. 125-144). Hoboken: John Wiley & Sons.

Fields, J., & Callegaro, M. (2007, December). *Background and Planning for Incorporating an Event History Calendar into the Re-Engineered SIPP*. Federal Committee on Statistical Methodology Meeting,

Freedman, D., Thornton, A., Camburn, D., Alwin, D., & Young-DeMarco, L. (1988). The life history calendar: A technique for collecting retrospective data. *Sociological Methodology, 18*(1), 37-68.

Glasner, T. (2011). *Reconstructing event histories in standardized survey research: Cognitive mechanisms and aided recall techniques*. Oisterwijk: Uitgeverij BOXPress.

Glasner, T., & van der Vaart, W. (2009). Applications of calendar instruments in social surveys: a review. *Quality & Quantity, 43*(3), 333-349. doi:10.1007/s11135-007-9129-8

Hoogendoorn, A. (2004). A Questionnaire Design for Dependent Interviewing that Addresses the Problem of Cognitive Satisficing. *Journal of Official Statistics, 20*(2), 219-232. Retrieved from http://www.jos.nu/Articles/abstract.asp?article=202219

Huinink, J., Brüderl, J., Nauck, B., Walper, S., Castiglioni, L., & Feldhaus, M. (2011). Panel Analysis of Intimate Relationships and Family Dynamics (pairfam): Conceptual framework and design. *Zeitschrift für Familienforschung, 23*(1), 77-101.

Jäckle, A. (2008). *The causes of seam effects in panel surveys* (No. 2008-14).

Jäckle, A. (2009). Dependent interviewing: A framework and application to current research. In P. Lynn (Ed.), *Methodology of Longitudinal Surveys* (pp. 93-111). Chichester: Wiley.

Jäckle, A., & Eckman, S. (2016). *Is that still the same? Has that changed? On the accuracy of measuring change with dependent interviewing* (Understanding Society Working Paper Series No. 2016-06). Colchester.

Jäckle, A., & Lynn, P. (2007). Dependent interviewing and seam effects in work history data. *Journal of Official Statistics, 23*(4), 529.

Kominski, R. (1990). *The SIPP event history calendar: Aiding respondents in the dating of longitudinal events* (Survey of Income and Program Participation Working Paper Series No. 132).

Krosnick, J. A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology, 5*(3), 213-236. doi:10.1002/acp.2350050305

Krosnick, J. A., Narayan, S., & Smith, W. R. (1996). Satisficing in surveys: Initial evidence. *New Directions for Evaluation, 1996*(70), 29-44. doi:10.1002/ev.1033

Lynn, P., Jäckle, A., Jenkins, S. P., & Sala, E. (2012). The impact of questioning method on measurement error in panel survey measures of benefit receipt: evidence from a validation study. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 175*(1), 289–308. doi:10.1111/j.1467-985X.2011.00717.x

Lynn, P., & Sala, E. (2006). Measuring Change in Employment Characteristics: The Effects of Dependent Interviewing. *International Journal of Public Opinion Research, 18*(4), 500–509. doi:10.1093/ijpor/edl013

Moore, J. (2012). *Analysis of Recorded Interviews in the 2010 SIPP-EHC Field Test* (SEHSD Working Paper No. 2012-17).

Moore, J., Bates, N., Pascale, J., & Okon, A. (2009). Tackling seam bias through questionnaire design. In P. Lynn (Ed.), *Methodology of Longitudinal Surveys* (pp. 73-92). Chichester: Wiley.

pairfam. (2015). *The German Family Panel (pairfam), Codebook Anchor, Wave 6, 2013/2014: Release 6.0.*

Pascale, J. (2009). *Event History Calendar Field Test Field Representative Focus Group Report* (Statistical Research Division Study Series (Survey Methodology) No. 02). Washington, D.C.

Reimer, M., & Matthes, B. (2007). Collecting Event Histories with TrueTales: Techniques to Improve Autobiographical Recall Problems in Standardized Interviews. *Quality & Quantity, 41*(5), 711-735. doi:10.1007/s11135-006-9021-y

Rips, L. J., Conrad, F. G., & Fricker, S. S. (2003). Straightening the Seam Effect in Panel Surveys. *Public Opinion Quarterly, 67*(4), 522-554. doi:10.1086/378962

Sayles, H., Belli, R. F., & Serrano, E. (2010). Interviewer Variance Between Event History Calendar and Conventional Questionnaire Interviews. *Public Opinion Quarterly, 74*(1), 140–153. doi:10.1093/poq/nfp089

Sudman, S., & Bradburn, N. M. (1974). *Response effects in surveys*: Chicago: Aldine.

Trahms, A., Matthes, B., & Ruland, M. (2016). Collecting Life-Course Data in a Panel Design: Why and How We Use Proactive Dependent Interviewing. In H.-P. Blossfeld, J. von Maurice, M. Bayer, & J. Skopek (Eds.), *Methodological Issues of Longitudinal Surveys: The Example of the National Educational Panel Study* (pp. 349-366). Wiesbaden: Springer Fachmedien Wiesbaden. Retrieved from http://dx.doi.org/10.1007/978-3-658-11994-2_20

Uhrig, N., & Sala, E. (2011). When Change Matters: An Analysis of Survey Interaction in Dependent Interviewing on the British Household Panel Study. Sociological Methods & Research, 40(2), 333-366.

van der Vaart, W. (2004). The time-line as a device to enhance recall in standardized research interviews: A split ballot study. *Journal of Official Statistics, 20*(2), 301-318.

Watson, N. (2009, July). *Disentangling Overlapping Seams: The experience of the HILDA Survey*. HILDA Survey Research Conference, University of Melbourne.

methods, data, analyses | Vol. 11(1), 2017, pp. 67-86 DOI: 10.12758/mda.2016.009

Gender of Interviewer Effects in a Multi-topic Centralized CATI Panel Survey

Oliver Lipps & Georg Lutz

Swiss Centre of Expertise in the Social Sciences (FORS)

Abstract

This paper is motivated by two recent articles which show that numerous studies which analyzed gender of interviewer effects did not take interviewer nonresponse selection effects into account. For example, interviewers may be more successful at recruiting respondents with characteristics similar to themselves and who give answers that are similar to their own, and this may result in spurious gender of interviewer effects. Our research is novel because it uses data from a large panel survey in which the same respondent is asked the same questions repeatedly by interviewers of random genders using the centralized telephone mode. We use the panel design to show the importance of checking for all relevant variables in models where selection may cause bias. To this end, we use respondent fixed effects models as a reference to yield unbiased coefficients.

We find gender of interviewer effects that are in line with social desirability theory on gender issues such as female discrimination. However, not all gender-related questions are affected by gender of interviewer effects and, in addition, we do not find any effects on political and (factual) household task related questions. In line with the notion of social distance, there is a higher likelihood that answers respondents are less comfortable with are given to interviewers of the same gender regarding (sensitive) health questions.

Keywords: Research of gender of interviewer effects; selection effects; social desirability; social role; social distance; fixed effects

Introduction

Gender of interviewer effects may cause severe answer bias (Groves et al., 1992; Davis et al., 2010). For example, in interviewer-based surveys, people may give more liberal answers to questions on women's rights to female interviewers than to male interviewers due to a wish to ensure a good atmosphere during the interview by providing answers that are assumed to be preferred by the interviewer. In addition, the nature of the answers given may depend on the match in characteristics between interviewer and respondent (e.g., Catania et al., 1996). Measurement errors may not be the only source of gender of interviewer effects, as even if interviewers are *assigned* to respondents at random, female interviewers may *interview* different sample members than male interviewers so that their respondent sample is different (Groves & Couper, 1998). Two recent articles addressed this issue: one focused on telephone surveys (West & Olson, 2010) and the other on face-to-face surveys (West et al., 2013). Each found that large parts of the interviewer variance were actually due to nonresponse error variance in addition to measurement error variance. However, to distinguish these error sources in cross-sectional surveys which are typically analyzed (West & Olson, 2010) is very difficult.

Compared with existing studies on gender of interviewer effects (see Table 2.2 in Davis, 2008), to the best of our knowledge our approach is the first to use a panel survey with a random assignment of interviewers of both sexes to respondents across waves to study gender of interviewer effects. In this model, the same respondent answers the same questions repeatedly, sometimes to male, sometimes to female interviewers, a design that guarantees that interviewers of both genders interview the same sample so that there is no gender of interviewer nonresponse effect. In addition, we use a large sample that is representative of a national residential population, and a large number of socio-demographically heterogeneous interviewers. This means the study is not a simple experimental low N-study with a highly selective use of respondents.

This article is organized as follows: after providing theoretical reasons for gender of interviewer effects and reviewing empirical findings, we describe the data and the survey design we use to analyze gender of interviewer effects and formulate the following hypotheses. We expect no gender of interviewer effects in domains where gender roles are unimportant, even if men and women give different answers. On the contrary, we expect more traditional answers to questions from male interviewers in domains where gender roles prevail. In addition, we expect more valid or honest answers to questions given by interviewers of the same gender

Direct correspondence to
Oliver Lipps, FORS, c/o University of Lausanne, 1015 Lausanne
E-mail: oliver.lipps@fors.unil.ch

on sensitive or embarrassing topics with which the respondent is less comfortable with. After introducing the models used, we present and discuss the results and offer our conclusion.

1 Theory and Empirical Findings

There are different reasons why people give different answers to male or female interviewers (Atkin & Chaffee, 1972; Cosper, 1972; Fowler & Mangione, 1990). According to *social desirability* theory (DeMaio, 1984; Paulhus, 2002), respondents reflect on what might be considered the mainstream views in a society on a given topic and then adapt their answer to this view. The likelihood of respondents giving a response which they think is more accepted by society may depend on interviewer characteristics, such as their gender. That men and women hold systematically different attitudes on a wide range of issues is widely known: men are typically more traditional and women are more liberal and more in favor of social welfare programs and equal rights (Blekesaune & Quadagno, 2003; Eagly & Steffen, 1984; Eagly et al., 2004; Pratto et al., 1997).

Differences are, however, not likely to be enough to produce interviewer effects, because most attitudes do not have a clear gender dimension. Some have argued that we should find interviewer effects only on issues that are based on *social role* theory (Diekman & Schneider, 2010). Social role theory asserts that interviewer gender effects occur when attitudes are linked to expectations about gender roles and gender equality. Gender stereotypes and expectations of gender roles are still widely present in western societies despite some signs of a decline (Wilde & Diekman, 2005), and Diekman & Goodfriend (2006) state that women still typically occupy social roles of care takers for others while men are assumed to take the role of leadership and power. Another theory states that communication is more comfortable over a smaller social distance (Groves et al., 1992; Liu & Stainback, 2013; Snell Dohrenwend et al., 1968; Tu & Liao, 2007). Respondents, when answering sensitive questions, may feel more at ease with interviewers with whom they have something in common, including the same gender. As a consequence, respondents give more valid or honest responses to such questions.

Previous research has focused a lot on interviewer effects on gender issues where the general expectation is that female interviewers are either more likely to produce more feminist or liberal responses (Lueptow et al., 1990) or, the other way around, male interviewers ellicit responses that "appear more traditional" (Flores-Macias & Lawson, 2008, p.100). Huddy et al. (1997), for example, find that respondents were more likely to give a feminist response to a female interviewer in two local-area telephone surveys in the U.S. on questions related to the women's movement, women's issues, and gender equality. Interviewers might affect men and women to a different extent, depending on the sensitivity of the question in the cultural context (Becker et al., 1995; Benstead, 2013). For example, Flores-Macias & Lawson (2008) find that interviewer gender is more likely to affect men living in (rather liberal) Mexico City than women regarding gender-sensitive questions, while Lueptow et al. (1990) find that male interviewers have more influence on the response variance of women in a Midwestern metropolitan area. In general, however, gender of interviewer effects are rather weak and sometimes inconsistent (see the review in Davis et al., 2010), especially concerning interviewer-respondent gender interaction effects. For example, Fuchs (2009) finds both opposite-gender and same-gender effects in a German CASI experiment, which contradicts social distance theory. And although Liu and Stainback (2013) find gender of interviewer effects on questions regarding the happiness of married persons compared to unmarried persons in a Chinese survey, they do not find differences according to respondents' gender.

Recently some scholars have investigated whether interviewer-specific nonresponse bias causes significant portions of "gender of interviewer" effects. Using factual questions, West and his colleagues analyzed how much of the interviewer effect is due to measurement error and how much is due to selection error. While West and Olson (2010) found substantial selection effects in a cross-sectional telephone survey, West et al. (2013) report selection effects in a cross-sectional face-to-face survey. The surveys used in the two articles were matched with administration data which contained the "true" values. Without the availability of such auxiliary data, the identification of the two interviewer error variances is not possible. In addition, interviewer effects can only be examined for factual variables, not attitudes. However, the latter may be more fruitful when analyzing gender of interviewer effects.

In the light of these inconclusive findings and weak data sources, Davis et al. (2010) argue for more research which uses designs in which respondents are randomly assigned to interviewers (Gillikin, 2008), and which utilze a large number of interviewers. Davis et al. (2010) complain that "[t]hese ideal study qualities may be difficult to achieve [...] However, even if lacking perfect design, the repeated investigation and reporting of interviewer effects, whether significant or null, will contribute to a significantly enhanced understanding of the magnitude and frequency of interviewer effects" (p. 24). In addition Davis (2008) calls for more telephone-administered studies (p. 28, 29). We have thus accommodated this by measuring gender of interviewer effects in different topic domains and done so more accurately by using panel data as it seems particularly suitable for this purpose. Specifically we test the following hypotheses:

H1: We expect gender of interviewer effects when two conditions are met: questions 1) show different answers between men and women and 2) relate to clear gender specific social roles. The latter means that there needs to be a gender specific dimension such as female rights or discrimination. Questions on more general topics such as general political questions are not sufficient even if different attitudes between the genders are present.

H2: In terms of the direction of effects, we expect that female interviewers prompt more liberal views, while male interviewers prompt more traditional opinions from respondents of both genders. We do not expect gender of interviewer-respondent matching effects

H3: We expect gender of interviewer-respondent matching effects on questions the answers to which may be embarrassing for respondents, even if they are not related to gender specific social roles. We expect more valid answers if the interviewer and the respondent have the same gender.

To test these hypotheses, we used data from Switzerland. Though gender equality is a constitutional norm and legislation to prohibit gender discrimination came into force in 1995 (Federal Authorities, 2013), some people still hold the view that women should play a more important role in the home while men should be the primary earner (Bernardi et al., 2013; Makarova & Herzog, 2015). Although female labor force participation is increasing (SFSO, 2015), this expansion is evident predominantly through part-time jobs, especially in lower-pay sectors with less responsibility (Bernardi et al., 2013), and wage differences are still substantial (Murphy & Oesch, 2015). While this has contributed to more heterogeneous life trajectories for women, men's life trajectories still correspond to the classical breadwinner model (Widmer et al., 2003). To explore if gender of interviewer effects are limited to gender related issues or whether this is a broader phenomenon, we included answers to questions in additional domains: politics (Huddy et al., 1997; Hutchinson & Wegge, 1991; Lipps & Lutz, 2010), the role of performing different household tasks (Ballou & DelBoca, 1980; Grimes & Hansen, 1984; Kane & Macauley, 1993, Klein & Kühhirt, 2010), and health (Davis et al., 2010).

2 Data

Telephone surveys are well suited to the study of interviewer gender effects (Davis & Silver, 2003; Grimes & Hansen, 1984; Groves & Magilavy, 1986; Huddy et al., 1997; Kane & Macaulay, 1993; Lueptow et al., 1990). While respondents are able to make an accurate guess about an interviewer's gender in a telephone interview (Callegaro et al., 2005), possibly distracting information about an interviewer's socioeconomic status, physical attractiveness, dress, personal demeanor, or other cues that might influence face-to-face survey responses are absent (Adenskaya & Dommeyer, 2011; Groves & Fultz, 1985). In addition, telephone surveys conducted from a telephone center use random interviewer-respondent assignments, thus reducing the risk of confusing area effects and interviewer effects as is often the case in face-to-face surveys (O'Muircheartaigh & Campanelli, 1998; West et al., 2013). Roberts et al. (2006) find that telephone respondents are more likely to give socially desirable responses than face-to-face respondents. Davis (1997), however, argues that telephone surveys should produce smaller effects because of the greater social distance inherent in using a phone line (also Fowler & Mangione, 1990).

We used data from the Swiss Household Panel (SHP; Voorpostel et al., 2015) which is an annual, centrally conducted and nationwide CATI panel survey, using a stratified random sample of the Swiss residential population. Starting in 1999 with more than 5,000 households, the SHP added two refreshment samples, one in 2004 with more than 2,500 households, and one in 2013 with about 4,000 households. In their respective first waves, the 1999 original sample household level response rate amounted to 64%, that of the 2004 refreshment sample was 65%, and that of the 2013 refreshment sample was 60% (RR1; AAPOR, 2015). Fieldwork is conducted each year between September and January using about 100 interviewers, and each year, the household reference person is asked to first complete the household grid questionnaire and then the household questionnaire, which includes among other questions the share of household tasks between the partners of a household. Finally, all household members aged 14 or over are interviewed using the individual questionnaire. The SHP contains a wide range of questions about health, well-being, attitudes, social networks and economics. Gender of interviewer information is available for almost all interviewers in 2000, and from 2003 on. Since not all questions investigated have been asked in all the years (2000, 2003-2014), the sample size is different according to the question analyzed. Interviews from 18,555 respondents interviewed by 605 interviewers with given gender are used. While about a third of the respondents report, respectively, one, two to four, and five to 13 waves, 65% of the interviewers work only one wave, 20% two waves, and about 15% work more than two waves. A third of the 605 interviewers are men. To rule out selection effects due to a different response rate, we analyzed response rate differences between male and female interviewers and ran two cluster robust logit models, using pooled data from the first contacts on the household grid level and on the individual questionnaire level, respectively. First contacts are crucial determinants of final cooperation and are well suited to investigate interviewer performance in centralized CATI surveys (Lipps, 2009), although about 60% of the households and individuals needed more than one contact to be finalized. After checking the survey year, whether the first contact occurred during the normal or the refusal conversion[1] field phase, the number of contacts, and the number of unsuccessful calls, results show that the predicted cooperation probabilities of male (female) interviewers amount to 82.1% (81.6%) on the household grid level and to 78.7% (78.9%)

[1] All households/individuals, who uttered a soft refusal during the normal field phase, were tried to be converted.

on the individual questionnaire level respectively. Both gender of interviewer differences are insignificant on the 1% level.

We selected questions where we found clear differences in the response behavior of men and women. Table 1 gives an overview of the different questions, the number of observations, and the mean values of men and women in relation to these questions. Four questions are gender specific, five questions are related to general politics attitudes, five questions ask about the distribution of household tasks, and five questions about health issues. The question wording and their exact answer categories are listed in the appendix.

Answers to all questions show highly significant differences between men and women, and the differences mostly go in the expected direction. This is the case for gender related measures: women believe more often that they are penalized, that there should be more measures to support women, that having a job is the best guarantee for women and men to be independent, and they disagree more often that pre-school children suffer when the mother works for pay. For example, men report 0.45 units less discrimination than women (first question). In terms of political questions, women are less in favor of a strong army than men; more in favor of environmental protection rather than economic growth, more against nuclear energy, and more in favor of increasing social expenses. The only unexpected result is that women are less often in favor of equality between Swiss and foreigners and hence more discriminatory than men. As for household tasks, women do more cleaning and washing, while men manage the finances and administrative tasks slightly more often. Not surprisingly, women report many more hours of housework. As for health issues, women tend to report more physical and mental health problems than men.

Table 1 Means of dependent variables by sex and t-tests of differences.

	N(obs.)	Women	Men	P(|T| >| t|)
GENDER				
Women in Switzerland are sometimes penalized (0=no-10=yes)	52212	5.41	4.96	0.000
There should be more measures to support women in Switzerland (0=no-10=yes)	51776	6.00	5.34	0.000
To have a job is the best guarantee for women and men to be independent (0=no-10=yes)	49821	8.26	8.01	0.000
A pre-school child suffers, if his or her mother works for pay (0=no-10=yes)	49205	5.25	6.18	0.000
POLITICS				
In favor of a strong Swiss army (0=no, 1=neither nor, 2=yes)	47802	0.97	1.08	0.000
Foreigners should have the same opportunities as Swiss (0=no, 1=neither nor, 2= yes)	51171	1.32	1.41	0.000
Environment should be more important than economic growth (0=no, 1=neither nor, 2=yes)	51248	1.41	1.25	0.000
Switzerland should continue to have nuclear energy (0=no, 1=neither nor, 2=yes)	50315	0.52	0.87	0.000
Switzerland should increase social expenses (0=no, 1=neither nor, 2=yes)	49465	1.25	1.13	0.000
HOUSEHOLD TASKS				
In our Household it is mostly me who does the cleaning (0=no, 1=yes)	43009	0.79	0.29	0.000
In our Household it is mostly me who does the laundry (0=no, 1=yes)	43009	0.88	0.25	0.000
In our Household it is mostly me who manages the finances (0=no, 1=yes)	43009	0.30	0.35	0.000
In our Household it is mostly me who handles administration (0=no, 1=yes)	43009	0.62	0.66	0.000
Hours of housework (per week)	70561	14.71	5.87	0.000
HEALTH				
Body weight (kg)	60405	63.91	78.42	0.000
Suffering from headaches during the past four weeks (0=no, 1=yes)	60933	0.40	0.27	0.000
Physical health bad (1=very good, …, 5=not well at all)	72414	2.00	1.89	0.000
Having the blues (0=never-10=always)	72371	2.35	1.69	0.000
Sadness (0=never-10=always)	48109	3.73	3.05	0.000

Data: Swiss Household Panel 2000, 2003-2014.

3 Variables and Modeling

Our independent *research* variables are interviewer gender and interviewer gender interacted with respondent gender. We used female as a reference gender category. A significant interviewer male coefficient b means that there is a difference by b between survey answers to a female and a male interviewer. A significant interaction coefficient b' (interviewer and respondent male match) means that a man, when interviewed by a man, exhibits a difference by b' to the situation, when interviewed by a woman (the main effects are controlled).

Our target is to estimate random effects (RE) models the gender of interviewer coefficients of which are close to those of the respective fixed effects (FE) models. Only then can we be sure that we have included the relevant respondent *time invariant* variables (Morgan & Winship, 2014) which we suspect are responsible for a great deal of possible selection effects. For example, the "naïve" mean difference to the first question between male and female interviewers amounts to 0.43 units, which is almost as high as the differences between male and female respondents (0.45) and which suggests that despite our design there are some selection effects. We aim to control for all variables which may have effects on the gender of interviewer sample selection, due to a selective accessibility and/or a selective cooperation (Groves & Couper, 1998). To decide which variables to include as controls for gender of interviewer sample selection effects, we tested mean differences between the samples of male and female interviewers for the following variables by means of T-tests. Respondent gender itself is not affected by gender of interviewer sample selection (P(|T| >|t|) = .257 (.053) on the household (person) level (but included in the RE model).

- Respondent is the household reference person (P(|T| >|t|) = .000)
- Household needed refusal conversion (P(|T| >|t|) = .026)
- Language region: Swiss-German speaking part (reference), French speaking part (P(|T| >|t|) = .000), Italian speaking part (P(|T| >|t|) = .000)
- Respondent's highest education (P(|T| >|t|) = .115)
- Respondent has a partner (P(|T| >|t|) = .077)
- Respondent is employed (P(|T| >|t|) = .000)
- Respondent lives in a city (P(|T| >|t|) = .071)
- Day of first contact: Monday (P(|T| >|t|) = .019), Tuesday (P(|T| >|t|) = .494), Wednesday (P(|T| >|t|) = .127), Thursday (P(|T| >|t|) = .580), Friday (P(|T| >|t|) = .073), Saturday (P(|T| >|t|) = .064).
- Time of first contact: before 2 pm (P(|T| >|t|) = .003), between 2 pm and 6 pm: (P(|T| >|t|) = .164), between 6 pm and 8 pm: (P(|T| >|t|) = .324), after 8 pm: (P(|T| >|t|) = .249)
- Number of contacts (P(|T| >|t|) = .249).

- Number of unsuccessful calls (calls with no contact) (P(|T| >|t|) = .031).
- Age of youngest child in the household: no child (reference), between 0 and 6 years: (P(|T| >|t|) = .029), between 7 and 17 years: (P(|T| >|t|) = .324)
- Survey wave to control for panel conditioning (Warren & Halpern-Manners, 2012): first (reference), second (P(|T| >|t|) = .428), third or higher (P(|T| >|t|) = .000)
- Survey year to account for time effects and for different interviewer compositions across different years: 2000 (reference), 2003, 2004, 2005, 2007, 2009, 2011-2012, 2014 (P(|T| >|t|) = .000), 2006 (P(|T| >|t|) = .013), 2008 (P(|T| >|t|) = .039), 2010 (P(|T| >|t|) = .014), 2013 (P(|T| >|t|) = .46)
- Age group: 14-25 years (reference), 26-35 years (P(|T| >|t|) = .035), 36-45 years (P(|T| >|t|) = .000), 46-55 years (P(|T| >|t|) = .000), 56-65 years (P(|T| >|t|) = .287), 66+ years (P(|T| >|t|) = .251)

Based on these findings we decided to control for all variables apart from education, partner, living in a city, age of youngest child in the household, day and time of first contact, number of contacts or unsuccessful calls, and whether the household needed refusal conversion (the latter five variables come from the CATI call data and are available only from 2005 on). Before we dropped the call data variables, we tested their joint significance in a linear regression of the residual of the dependent variables (see Table 1) on all other variables. As it turns out, these call data variables have little additional explanatory power. Exceptions are significant (1%) F-values in two gender models, two household tasks models, and two health models. However, all models have a McFadden Pseudo R^2 smaller than .003. Using the research and the control variables, we ran Hausman tests to test the differences between the coefficients of the (consistent but less efficient) FE model and the coefficients of the (more efficient) RE model.

Finally, we tested whether there are gender of interviewer effects depending on other respondent characteristics. For example Huddy et al. (1997) found that that gender of interviewer effects were more pronounced among younger respondents. We found that gender of interviewer in interaction with respondent age groups did not show substantial effects, with the exception of younger people who are less gender of interviewer sensitive when asked whether women are discriminated against, and – surprisingly – young people who report a higher weight to male interviewers. In the end, we decided not to include interactions of respondent socio-demographic characteristics other than gender with interviewer gender.

4 Results

The Hausman tests of all models are significant on the 1% level. This means that for all issues there are time invariant respondent specific omitted variables which cause biased gender of interviewer effects in RE models. Nevertheless, the size of the coefficients of the FE gender of the interviewer are very similar to those of the RE gender of the interviewer for most dependent variables (see Table 2). As an example of our success in controlling for some relevant gender of interviewer selection variables, the mean difference between male and female interviewers to answers on the first question is now reduced to about 0.25 units. Interaction gender of the respondent * gender of the interviewer coefficients show a greater difference between the FE and the RE models. Since the FE models yield consistent parameter estimates (Morgan & Winship, 2014) we rely on these modeling results to interpret the gender of interviewer effects even if we lose some statistical precision. For the sake of completeness, however, we list both the FE and the RE coefficients. Since all Hausman tests are significant on the 1% level we do not list the respective significance separately.

Table 2 Random and fixed linear effects model coefficients of interviewer male and interaction interviewer male*respondent male.

Coefficient of Gender of Interviewer male and interaction interviewer male*respondent male	N	RE-models		FE-models	
		Iwer male	both male	Iwer male	both male
GENDER					
Women in Switzerland are sometimes penalized (0=no-10=yes)	52,210	-.254*	-.092$^+$	-.255*	-.063$^+$
There should be more measures to support women in Switzerland (0=no-10=yes)	51,774	-.382*	-.088$^+$	-.370*	-.077
To have a job is the best guarantee for women and men to be independent (0=no-10=yes)	49,819	-.109*	.039	-.119*	.034
A pre-school child suffers, if his or her mother works for pay (0=no-10=yes)	49,203	.011	-.029	.013	-.048
POLITICS					
In favor of a strong Swiss army (0=no, 1=neither nor, 2=yes)	47,800	-.006	-.010	-.010	-.004
Foreigners should have the same opportunities as Swiss (0=no, 1=neither nor, 2= yes)	51,169	.010	-.007	.010	-.004
Environment should be more important than economic growth (0=no, 1=neither nor, 2=yes)	51,246	-.012	.007	-.009	.008

Coefficient of Gender of Interviewer male and interaction interviewer male*respondent male	N	RE-models		FE-models	
		Iwer male	both male	Iwer male	both male
Switzerland should continue to have nuclear energy (0=no, 1=neither nor, 2=yes)	50,313	-.002	-.015	-.009	-.013
Switzerland should increase social expenses (0=no, 1=neither nor, 2=yes)	49,463	-.032*	.005	-.033*	.007
HOUSEHOLD TASKS					
In our Household it is mostly me who does the cleaning (0=no-1=yes)	43,008	-.000	-.007	-.000	-.006
In our Household it is mostly me who does the laundry (0=no-1=yes)	43,008	-.002	.005	-.001	.005
In our Household it is mostly me who manages the finances (0=no-1=yes)	43,008	.002	.004	.001	.009
In our Household it is mostly me who does administration (0=no-1=yes)	43,008	.001	.001	.000	.006
Hours of housework (per week)	70,559	-.149$^+$	.183	-.132	.167
HEALTH					
Body weight (kg)	60,403	-.264*	.434*	-.279*	.452*
Suffering from headaches during the past four weeks	60,931	-.017*	.019*	-.019*	.020*
Physical health bad (1=very good, …, 5=not well at all)	72,412	-.037*	.016	-.035*	.011
Having the blues (0=never-10=always)	72,369	-.103*	.063$^+$	-.111*	.071*
Sadness (0=never-10=always)	48,109	-.089*	.018	-.091*	.007

Data: Swiss Household Panel 2000, 2003-2014. Mean number of observations by respondent between 3.287 (Sadness) and 3.904 (bad health). Models controlled for reference person (all but Household Tasks), respondent male (RE Models), language region, being employed, age, respondent wave (first, second, third or later), year dummies. *=significant on 1% level += significant on 5% level.

In the following we briefly describe the gender of interviewer effects across the different domains and discuss if our hypotheses are confirmed or not.

Gender

When asked by a man, both genders advocate more traditional positions towards women's discrimination and measures to support women, as well as about the independence which a job guarantees. Interestingly, the gender of interviewer effect on these three questions is as high as about half of the gender of the respondent effect (see table 1). Generally, there are weak or no respondent-interviewer gender match effects. We find no gender of interviewer effects on the question whether a pre-school child suffers if his or her mother works for pay. An explanation for the lack of interviewer effects for this question about child suffering could be that unlike the first three questions in this domain, the job related question for a mother with a small child is interesting only for those concerned. We run the model about child suffering for the sample of people who work and have a child under the age of 7 years at home. For this sample, we find a positive interviewer=male effect for women and a zero effect for men. It is possible that the three (more general) initial questions are less deeply reflected upon, and more socially desired answers are mechanically provided in response to these questions.

Politics

With one exception ("Switzerland should increase social expenses"), items on political attitudes are not affected by the interviewer gender.

Household Tasks

There are no gender of interviewer effects on household task items.

Health

For health issues, we find evidence that men, and even more so women, report better physical and mental health when interviewed by a man. This shows that health issues are potentially sensitive and women are trusted more. Interestingly, both genders, and especially women, report lower body weight when interviewed by an interviewer of the opposite sex: women report .279 kg less, men a significant (1%) .452 kg - .279 kg = .173 kg less. Reporting a high body weight to an interviewer of the opposite sex may be embarrassing. The same holds for headaches and in parts for having the blues, where interviewers of the same sex "admit" worse health conditions. For example, in terms of headaches, the main effect of -.019 from male interviewers means that women report 1.9 % points less occurrences to male than to female interviewers, while men report the same amounts to interviewers of both genders (-.019+.020 is insignificant). Similarly, while women report .111 less often having the blues to male interviewers, men again show no gender of interviewer effect (-111+.071 is insignificant). Male interviewers receive reports of better physical health and less levels of sadness by respondents of both genders.

Before concluding, we compare our findings with our hypotheses:

H1 (effects if answers differ between men and women and if questions have a gender specific social role): We find that answers on *general* gender issues are affected by the gender of the interviewer. Questions from which respondents are not concerned, are not affected. Political questions and household tasks are not affected. Hypothesis H1 is confirmed to a great extent.

H2 (male interviewers produce more traditional answers and interaction effects): When significant, the sign of the interviewer=male effect is in the expected direction. Interaction effects occur only on health questions. Hypothesis H2 is in parts confirmed.

H3 (gender of interviewer effects on sensitive questions if the interviewer and the respondent have the same gender): We find gender of interviewer effects on health issues, especially for female respondents: Women in particular feel embarrassed to disclose bad mental or physical health to men, although there is not a social role involved with health issues. Men, as representing the "strong sex", are reported better health, by both genders. For some issues, interviewers of the same gender are trusted more. This shows that for sensitive questions like health, social distance may play a role. Our hypothesis H3 is partly confirmed.

5 Conclusion

The motivation for this article comes from two recent articles which show that omitted interviewer nonresponse selection effects may have resulted in spurious gender of interviewer effects in a number of studies, even if interviewers are assigned to *sample members* at random. To examine gender of interviewer effects, this article uses data from the Swiss Household Panel (SHP), a large centralized CATI panel with randomly assigned interviewers to respondents. The same survey questions are answered repeatedly by the same respondent to interviewers of both genders. When designing the models, we remained concerned about selection effects because even in the SHP, interviewers of a certain gender may select respondents with omitted characteristics. If these characteristics are correlated with the dependent variable, the gender of interviewer coefficient will be biased. FE models eliminate this error if the characteristics are time-invariant. A careful choice of control variables yielded very similar gender of interviewer estimates in RE and FE models. This makes us confident that we have included the relevant variables in the models.

We find the expected effects of gender of interviewer on *general* gender issues (such as on female discrimination in Switzerland) where social desirability and social role theory has a sufficient impact. Not all gender-related questions are affected by gender of interviewer effects, although there are differences between the respondents' genders. This is true for the statements about "A pre-school child suffers, if his or her mother works for pay". Our explanation for the lack of gender of interviewer effects on this question is that unlike the (rather general) other gender-related questions, only a small part of the sample is personally affected. Consequently we find gender of interviewer effects on this question for women with small children at home. We do not find effects on political and on (factual) household task related questions. (Sensitive) health questions are affected by the gender of the interviewer. In particular, women report better physical and mental health to male interviewers, while, on the contrary, respondents report a higher body weight to interviewers of the same gender. This points to social desirability effects and the impact of social role theory on gender-related questions and the theory of social distance on health questions.

For researchers who work with cross-sectional data and who like to estimate unbiased gender of interviewer effects, our research shows the importance of including all variables which have an effect on interviewer sample selection and the dependent variable. Suppose a researcher analyzes gender of interviewer effects on the female discrimination question (our first dependent variable). If the survey was conducted in a country with regions of different cultural contexts with different distributions of male and female interviewers and with different gender of interviewer effects, these contexts (in our case the language regions) must be controlled for. This sounds trivial but may be ignored by a number of researchers who are less familiar with the design of the survey at hand. Omitting the language region in a simple pooled OLS model would result in an interviewer=male coefficient of -.356, and this estimate would amount to -.260 if region is controlled. The latter comes very close to our estimate of -.254 in a RE model and -.255 in a FE model. Of course FE models are not without problems: FE models yield biased gender of interviewer coefficients if a changed interviewer gender goes with a parallel change of a related (unobserved) variable. For example, if men tend to interview at different times than women and these different times are correlated with attitudes of the then recruited respondents, the gender of the interviewer coefficient will be biased. However we believe that our random respondent-interviewer assignment and our control variables yield sufficiently unbiased gender of the interviewer coefficients for the variables analyzed.

Literature

Adenskaya, L., & Dommeyer, C. (2011). Can perfume increase the response rate to a face-to-face survey? *International Business and Economics Research Journal*, 3(2), 37-43.

Atkin, C.K., & Chaffee, S.H. (1972). Instrumental response strategies in opinion interviews. *Public Opinion Quarterly*, 36(1), 69-79. doi: 10.1086/267976

Ballou, J., & DelBoca, F.K. (1980). Gender interaction effects on survey measures in telephone interviews. Paper presented at the 35th annual meeting of the American Association of Public Opinion Research, Mason, Ohio.

Becker S., Feyisetan K., & Makinwa-Adebusoye P. (1995). The effect of the sex of interviewers on the quality of data in a Nigerian family planning questionnaire. *Studies in Family Planning*, 26(4): 233-240. Article stable URL: http://www.jstor.org/stable/2137848.

Benstead, L. (2013). Effects of interviewer-respondent gender interaction on attitudes toward women and politics: Findings from Morocco. *International Journal of Public Opinion Research* Online publication date: 27-Sep-2013. doi: 10.1093/ijpor/edt024

Bernardi, L., Ryser, V.-A., & Le Goff, J.-M. (2013). Gender role-set, family orientations, and women's fertility intentions in Switzerland. *Swiss Journal of Sociology*, 39(1), 9-31.

Blekesaune, M., & Quadagno, J. (2003). Public attitudes toward welfare state policies a comparative analysis of 24 nations. *European Sociological Review,* 19(5), 415-427. doi: 10.1093/esr/19.5.415

Callegaro, M., De Keulenaer, F. Krosnick, J.A., & Daves, R.P. (2005). Interviewer effects in a RDD telephone pre-election poll in Minneapolis 2001. An analysis of the effects of interviewer race and gender. *Proceedings of the American Statistical Association, 60th Annual Conference of the American Association for Public Opinion Research,* 3815-3821.

Catania, J.A., Binson, D., Canchola, J., Pollack, L.M., Hauck, W., & Coates, T.J. (1996). Effects of interviewer gender, interviewer choice, and item wording on responses to questions concerning sexual behavior. *Public Opinion Quarterly, 60*, 345-375. doi: 10.1086/297758

Cosper, R. (1972). Interviewer effect in a survey of drinking practices. *The Sociological Quarterly,* 13, 228-236. doi: 10.1111/j.1533-8525.1972.tb00806.x

Davis, D.W. (1997). Nonrandom measurement error and race of interviewer effects among African Americans. *Public Opinion Quarterly,* 61, 187-207.

Davis, D.W., & Silver, B.D. (2003). Stereotype threat and race of interviewer effects in a survey on political knowledge. *American Journal of Political Science,* 47(1), 33-45. doi: 10.1111/1540-5907.00003

Davis, R.E. (2008). "Whatever it means to you": Ethnicity, language, and the survey response in telephone administered health surveys of African Americans. PhD dissertation, University of Michigan.

Davis, R.E., Couper, M., Janz, N., Caldwell, C., & Resnicow, K. (2010). Interviewer effects in public health surveys. *Health Education Research,* 25(1), 14-26. doi: 10.1093/her/cyp046

DeMaio, T. (1984). Social desirability and survey measurement: A review. In C.F. Turner and E. Martin (Eds.), *Surveying Subjective Phenomena* (pp. 257-282), New York: Russell Sage.

Diekman, A.B., & Goodfriend, W. (2006). Rolling with the changes: A role congruity perspective on gender norms. *Psychology of Women Quarterly,* 30(4), 369-383. doi: 10.1111/j.1471-6402.2006.00312.x

Diekman, A.B. & Schneider, M.C. (2010). A social role theory perspective on gender gaps in political attitudes. *Psychology of Women Quarterly,* 34(4), 486-497. doi: 10.1111/j.1471-6402.2010.01598.x

Eagly, A.H., Diekman, A.B., Johannesen-Schmidt, M.C., & Koenig, A.M. (2004). Gender gaps in sociopolitical attitudes: a social psychological analysis. *Journal of Personality and Social Psychology*, 87(6), 796-816. doi: 10.1037/0022-3514.87.6.796

Eagly, A.H., & Steffen, V.J. (1984). Gender stereotypes stem from the distribution of women and men into social roles. *Journal of personality and Social Psychology,* 46(4), 735-754. doi: 10.1037/0022-3514.46.4.735

Federal Authorities (2013). 151.1 Federal Act of 24 March 1995 on Gender Equality (Gender Equality Act, GEA) (accessed 3NOV15: https://www.admin.ch/opc/en/classified-compilation/19950082/index.html.

Flores-Macias, F., & Lawson, C. (2008). Effects of interviewer gender on survey responses: Findings from a household survey in Mexico. *International Journal of Public Opinion Research,* 20(1), 100-110. doi: 10.1093/ijpor/edn007

Fowler, F., & Mangione, T. (1990). Standardized survey interviewing: minimizing interviewer-related error. Newbury Park: Sage.

Fuchs, M. (2009). Gender-of-interviewer effects in a video-enhanced web survey: Results from a randomized field experiment. *Social Psychology,* 40, 37-42. doi: 10.1027/1864-9335.40.1.37

Gillikin, J. (2008). Interpenetrated design. In P. Lavrakas (Ed.), *Encyclopedia of Survey Research Methods*, Sage. doi**:** http://dx.doi.org/10.4135/9781412963947

Grimes, M., & Hansen, G. (1984). Response bias in sex-role attitude measurement. *Sex Roles,* 10(1-2), 67-72. doi: 10.1007/BF00287747

Groves, R., Cialdini, R., & Couper, M. (1992) Understanding the decision to participate in a survey. *Public Opinion Quarterly,* 56(4), 475-495. doi: 10.1086/269338

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

Groves, R., & Fultz, N.H. (1985). Gender effects among telephone interviewers in a survey of economic attitudes. *Sociological Methods and Research,* 14, 31-52. doi: 10.1177/0049124185014001002

Groves R., & Magilavy, L. (1986). Measuring and explaining interviewer effects in centralized telephone surveys. *Public Opinion Quarterly,* 50, 251–266. doi: 10.1086/268979

Warren, J.R., & Halpern-Manners, A. (2012). Panel conditioning in longitudinal social science surveys. *Sociological Methods & Research*, 41(4), 491-534.

Huddy, L., Billig, J., Bracciodieta, J., Hoeffler, L., Moynihan, P.J., & Pugliani, P. (1997). The effect of interviewer gender on the survey response. *Political Behavior,* 19(3), 197-220. doi: 10.1023/A:1024882714254

Hutchinson, K.L., & Wegge, D.G. (1991). The effects of interviewer gender upon response in telephone survey-research. *Journal of Social Behavior and Personality,* 6(3), 573-84.

Kane, E.W., & Macaulay, L.J. (1993). Interviewer gender and gender attitudes. *Public Opinion Quarterly,* 57(1), 1-28. doi: 10.1086/269352

Klein, M., & Kühhirt, M. (2010). Sozial erwünschtes Antwortverhalten bezüglich der Teilung häuslicher Arbeit („Social desirability and Response Bias in case of the division of household labor“). *Methoden-daten-analysen,* 4(2), 79-104. PID: http://nbn-resolving.de/urn:nbn:de:0168-ssoar-210124.

Lipps, O. (2009). Cooperation in centralised CATI household panel surveys - a contact-based multilevel analysis to examine interviewer, respondent, and fieldwork process effects. *Journal of Official Statistics,* 25(3), 323-338.

Lipps, O., & Lutz, G. (2010). How answers on political attitudes are shaped by interviewers: Evidence from a panel survey. *Swiss Journal of Sociology,* 2, 345-358.

Liu, M., & Stainback, K. (2013). Interviewer gender effects on survey responses to marriage-related questions. Public Opinion Quarterly, nft019.

Lueptow, L.B., Moser, S.L., & Pendleton, B.F. (1990). Gender and response effects in telephone interviews about gender characteristics. *Sex Roles,* 22 (1/2), 29-42. doi: 10.1007/BF00288152

Makarova, E., & Herzog, W. (2015). Gender roles within the family: A study across three language regions of Switzerland. *Psychology of Gender through the Lens of Culture* (pp. 239-264). Springer International Publishing.

Morgan, S.L., & Winship, C. (2014). Counterfactuals and causal inference. Cambridge University Press.

Murphy, E., & Oesch, D. (2015). The feminization of occupations and change in wages: A panel analysis of Britain, Germany and Switzerland. *SOEP papers,* 731.

OECD (2006). Women and men in OECD countries. Accessed Jan 20, 2014: http://www.oecd.org/std/37962502.pdf.

O'Muircheartaigh, C., & Campanelli, P. (1998). The relative impact of interviewer effects and sample design effects on survey precision. *Journal of the Royal Statistical Society: Series A* (Statistics in Society), 161(1), 63-77.

Paulhus, D. (2002). Socially desirable responding: The evolution of a construct. In H. Braun, D. Jackson, & D. Wiley (Eds.), The role of constructs in psychological and educational measurement (pp. 49-69). Mahwah: Erlbaum.

Pratto, F., Stallworth, L.M., & Sidanius, J. (1997). The gender gap: Differences in political attitudes and social dominance orientation. *British Journal of Social Psychology,* 36(1), 49-68. doi: 10.1111/j.2044-8309.1997.tb01118.x

Roberts, C., Jäckle, A., & Lynn, P. (2006). Causes of mode effects: Separating out interviewer and stimulus effects in comparisons of face-to-face and telephone surveys. AAPOR - ASA Section on Survey Research Methods, 4221-4228.

SFSO (Swiss Federal Statistical Office) (2015). Gender equality – Economic activity of mothers. Accessed Oct 26, 2015: www.bfs.admin.ch/bfs/portal/en/index/themen/20/05/blank/key/Vereinbarkeit/01.html.

Snell Dohrenwend, B., Colombotos, J., & Dohrenwend, B. (1968). Social distance and interviewer effects. *Public Opinion Quarterly*, 32(3), 410-442.

Tu, S., & Liao, P. (2007). Social distance, respondent cooperation, and item nonresponse in sex survey. *Quality and Quantity,* 41, 177–199. doi: 10.1007/s11135-007-9088-0

Voorpostel, M., Tillmann, R., Lebert, F., Kuhn, U., Lipps, O., Ryser, V.-A., Schmid, F., Antal, E., Monsch, G.-A., & Wernli, B. (2015). Swiss Household Panel user guide (1999-2014), Wave 16, December 2015. Lausanne: FORS.

Wilde, A., & Diekman, A.B. (2005). Cross-cultural similarities and differences in dynamic stereotypes: A comparison between Germany and the United States. *Psychology of Women Quarterly,* 29(2), 188-196. doi: 10.1111/j.1471-6402.2005.00181.x

West, B., Kreuter, F., & Jaenichen, U. (2013). "Interviewer" effects in face-to-face surveys: A function of sampling, measurement error, or nonresponse? *Journal of Official Statistics,* 29(2), 277-297. doi: 10.2478/jos-2013-0023

West, B., & Olson, K. (2010). How much of interviewer variance is really nonresponse error variance? *Public Opinion Quarterly*, 74(5), 1004-1026.

Widmer, E., Levy, R., Pollien, A., Hammer, R., & Gauthier, J. A. (2003). Entre standardisation, individualisation et sexuation: une analyse des trajectoires personnelles en Suisse. *Swiss Journal of Sociology*, 29(1), 35-67.

Appendix

Dependent variable questions with number observations, means and standard deviations

The questions are asked in German, French or Italian, depending on the language of the respondent. Details of the wording in the different languages (including English), for example in wave 13, can be found at: http://www.swisspanel.ch/IMG/pdf/QuestionML-P-W13.pdf

Gender

- Do you have the feeling that in Switzerland women are penalized compared with men in certain areas? (0=not at all penalized, …, 10=strongly penalized)
- Are you in favor of Switzerland taking more steps to ensure the promotion of women? (0=not at all in favor, …, 10=totally in favor)
- To have a job is the best guarantee for a woman as for a man to be independent. (0=completely disagree, …, 10=completely agree)
- A pre-school child suffers, if his or her mother works for pay. (0=completely disagree, …, 10= completely agree)

Politics

- Are you in favor of Switzerland having a strong army or for Switzerland not having an army? (2=strong army, 1=neither nor, 0=no army)
- Are you in favor of Switzerland offering foreigners the same opportunities as those offered to Swiss citizens or in favor of Switzerland offering Swiss citizens better opportunities? (2=same opportunities, 1=undecided, 0=in favor of better opportunities for Swiss citizens)
- Are you in favor of Switzerland being more concerned with protection of the environment than with economic growth, or in favor of Switzerland being more concerned with economic growth than with protection of the environment? (2=in favor of stronger protection of the environment, 1=undecided, 0=in favor of stronger economic growth)
- Are you in favor of Switzerland having nuclear energy, or are you in favor of Switzerland not having nuclear energy? (2=in favor of Switzerland having nuclear energy, 1=undecided, 0=in favor of Switzerland not having nuclear energy)
- Are you in favor of a reduction or in favor of an increase of the Confederation's social spending? (2=in favor of an increase, 1=undecided, 0=in favor of leaving the same of a reduction)

Household Tasks

- Generally, who takes care of the cleaning or tiding up in your household? (1=mostly me, 0=another person)
- Generally, who takes care of the washing or ironing in your household? (1=mostly me, 0=another person)
- Generally, who manages the finances in your household? (1=mostly me, 0=another person)
- Generally, who does the administration in your household? (1=mostly me, 0=another person)
- On average, how many hours do you spend on housework (washing, cooking, cleaning) in a normal week?

Health

- How much do you weigh (in kg without clothes)?
- During the last 4 weeks, have you suffered from headaches? (0=not at all, 1=somewhat or very much)
- How is your health in general? (scale reversed to 4=very bad, 3=bad, 2=fair, 1=good, 0=very good)
- Do you often have negative feelings such as having the blues, being desperate, suffering from anxiety or depression? (0=never, …, 10=always)
- How frequently are you generally sad? (0=never, …, 10=always)

Determinants of Wealth Fluctuation: Changes in Hard-To-Measure Economic Variables in a Panel Study

Fabian T. Pfeffer & Jamie Griffin
University of Michigan

Abstract

Measuring fluctuation in families' economic conditions is the *raison d'être* of household panel studies. Accordingly, a particularly challenging critique is that *extreme* fluctuation in measured economic characteristics might indicate compounding measurement error rather than actual changes in families' economic wellbeing. In this article, we address this claim by moving beyond the assumption that particularly large fluctuation in economic conditions might be too large to be realistic. Instead, we examine predictors of large fluctuation, capturing sources related to actual socio-economic changes as well as potential sources of measurement error.

Using the Panel Study of Income Dynamics, we study between-wave changes in a dimension of economic wellbeing that is especially hard to measure, namely, net worth as an indicator of total family wealth. Our results demonstrate that even very large between-wave changes in net worth can be attributed to actual socio-economic and demographic processes. We do, however, also identify a potential source of measurement error that contributes to large wealth fluctuation, namely, the treatment of incomplete information, presenting a pervasive challenge for any longitudinal survey that includes questions on economic assets. Our results point to ways for improving wealth variables both in the data collection process (e.g., by measuring active savings) and in data processing (e.g., by improving imputation algorithms).

Keywords: wealth, panel study, active savings, measurement error, imputation

1 Motivation

Our understanding of families' economic wellbeing depends not only on how well we capture their current socio-economic conditions but also their movement within the economic hierarchy across time. In fact, the measurement of fluctuation in families' economic conditions could be considered the primary *raison d'être* of household panel studies (Duncan, 1984). In this research note, we reveal some of the factors that contribute to or jeopardize the ability of family household panel studies to accurately capture the changing economic fortunes of families. Doing so is particularly pressing in the context of an emerging new field of empirical inquiry: After decades of research on the dynamics of family income, recent scientific and public debate is increasingly focused on family wealth, or net worth, as a different and important dimension of economic wellbeing (e.g., Pfeffer & Schoeni, 2016; Piketty, 2013). The dynamics of wealth are of particular interest, for instance, to understand families' ability to smooth consumption during times of economic distress (Deaton, 1991) and to provide intergenerational support both in terms of investing in the young and caring for the elderly (Conley, 2001).

However, wealth information can be challenging to collect, and panel surveys that seek to measure its fluctuation over time face additional challenges (Bucks & Pence, 2015). In particular, a number of researchers have noted that wealth data tend to be noisier than many other economic data and have suggested that extreme fluctuation in wealth may result from measurement error (Bosworth & Smart, 2009; Hill, 2006; Venti, 2011). Here, we assume that even extreme wealth fluctuation is driven partly by real economic changes and partly by measurement error and our empirical analyses demonstrate the relative role of potential factors on both sides. Specifically, we consider households' demographic changes, economic behaviors and circumstances on the one hand, and two potential sources of measurement error on the other hand: "Observational errors" that might stem from a change in survey respondents and, more importantly, "errors of non-observation" (Groves, 2004) that arise from item nonresponse and its handling in the data processing phase.

We analyze data from the Panel Study of Income Dynamics (PSID, 2015), a study that not only has the distinguished record of being the world's longest-running nationally representative household panel study, but that also – and impor-

Acknowledgments

Funding for this research was provided by the National Institute on Aging (R01AG019802). The authors thank Vicki Freedman and Charlie Brown for helpful comments on an earlier version of this manuscript as well as the editors and reviewers.

Direct correspondence to

Fabian T. Pfeffer
Institute for Social Research, 426 Thompson Street, Ann Arbor, MI 48104, USA
E-mail: fpfeffer@umich.edu

tantly for this project – began to field a detailed survey module on families' assets in 1984. Our analyses identify some of the successes and limitations of the PSID asset module and, more generally, inform both data collection and data processing strategies for other household panel studies.

We begin by briefly summarizing some of PSID's main strategies for collecting, editing, and processing wealth data. We then describe our sample, main variables, and analytic approach. Our empirical findings address the distribution of between-wave changes in net worth and their predictors. We conclude by discussing the implications these findings have for the longitudinal collection of high-quality wealth information utilized by a rapidly growing field of empirical research.

2 Wealth Measurement in the PSID

The PSID started in 1968 and has collected a large set of socio-economic indicators for families and their descendants every year until 1997 and every other year since then. In 1984, it implemented a detailed module to measure families' assets. This module was repeated every five years until 1999 and every wave since then, amounting to a total of 12 waves of wealth data by 2015. The specific assets that form part of these data are listed later; here, we describe some of the strategies PSID employs during data collection, editing, and processing to reduce measurement error in its wealth variables. Many of these strategies were implemented in the first wealth survey of 1984 and were then state-of-the art. Some of these strategies still are; however, as we will show, others might be ripe to revisit given more recent methodological advances.

2.1 Data Collection: Unfolding Brackets

For each asset, respondents are first asked whether they own such asset (e.g., a home). For those who answer yes, the follow-up question asks about the value of the asset, sometimes with separate questions about the gross value (e.g., current home value) and the outstanding debt held against the asset (e.g., mortgages). To minimize the incidence of missing data in the collection of asset values, the PSID introduced a surveying technique that has become known as the "unfolding bracket" approach and that is now in use in a range of other major surveys (e.g., the Health and Retirement Study [HRS]). Respondents who report that they do not know an exact asset value receive a series of follow-up questions that ask them to report whether the value falls within certain pre-specified ranges ("brackets") (Juster & Smith, 1997). These brackets "unfold" as the interviewer asks for a dependent sequence of threshold values (e.g., "Does [X] amount to $10,000 or more?" If yes: "Does it amount to $50,000 or more?" If no: "Does it amount to $1,000 or more?"). In the PSID, this technique helps keep the prevalence of item nonresponse across a variety of different assets relatively low (reported below in Table 1). However, it also requires the assignment of a continuous value within those reported brackets, especially if individual asset components are to be cumulated to create a measure of total net worth.

2.2 Data Editing: Individual Lookups

In the data editing process, the PSID attempts to correct errors of observation that arise from either respondents or interviewers by investigating outlying responses and reconciling them with other information collected in the same or prior waves through individual lookups – a labor- and time-intensive process. Importantly, the outlying values for a given variable are defined only with respect to the distribution of that variable within the current survey wave. Conversely, other studies incorporate prior-wave information in the editing stage or even during data collection. For instance, the HRS preloads wealth values from the prior-wave interview and asks respondents to reconcile conflicting responses between the current and prior wave. In 2012, this procedure identified a small number of cases (≤ 2.5%) who corrected errors in either the prior or current wave.

2.3 Data Processing: Imputation

Finally, and most important for the purpose of this contribution, the PSID applies imputation procedures to fill in missing continuous asset values arising from item nonresponse and bracketed responses. Random hot-deck imputation procedures are used in the following sequence of steps (see also PSID, 2013, pp. 41-42): First, when a respondent does not report whether or not an asset (debt) is held, a yes or no value is randomly assigned with probabilities equal to the distribution of observed yes or no values. Second, for those reporting neither a continuous nor a bracket response for the value of the asset (debt), a bracket (e.g., $10,000 - $50,000) is randomly selected with selection probabilities equal to the distribution of observed brackets. Finally, all respondents who do not provide a continuous value for the asset (debt) (steps 1 and 2) are assigned a continuous value by randomly selecting an observed value within a given bracket and with selection probabilities equal to the distribution of observed continuous values within the respective bracket.[1]

Table 1 reports the share of cases with unknown continuous asset values for each wealth component, that is, those to which the described imputation procedure is applied (in years 2005 and 2007 for reasons described later). The extent of imputation differs substantially across wealth components (upper panel of Table 1), with

[1] The imputation approach differs somewhat for home equity as described in detail elsewhere (PSID, 2013, pp. 55-56).

Table 1 Item Nonresponse in Wealth Components and Net Worth
N=7,051

		Share of Item-Nonresponse (%)		
Wealth Component	Variable Names	2005	2007	Overall
Vehicles/motor homes/trucks/etc.	S713A / S813A	7.9	9.0	14.0
Checking/savings/money order/etc.	S705A / S805A	7.1	7.3	11.6
Retirement wealth (annuity/IRA)	S719A / S819A	4.9	4.3	7.7
Home equity (value-mortgages)	S720A / S820A	4.2	4.7	7.6
Stocks/mutual funds/etc.	S711A / S811A	4.7	4.6	7.6
Other financial assets (bond funds/estate/etc.)	S715A / S815A	4.1	3.9	7.2
Farm and business wealth	S703A / S803A	3.0	3.5	5.3
Other debt (credit card/student loans/etc.)	S707A / S807A	1.5	1.6	2.8
Other real estate	S709A / S809A	1.5	1.6	2.8
Across all components (= net worth measure)	(S717A / S817A)			
Average		4.3	4.5	7.4
Cumulative				
Zero		75.8	74.0	62.4
One		15.6	17.0	19.1
Two		4.7	5.5	8.5
Three or more		3.9	3.5	10.0

Note: The overall column reports the share of cases with a specific wealth component imputed in either 2005 or 2007 (or both) and the total number of components missing across both years; N=7,051

the largest share of cases requiring imputation for the continuous value of vehicles (8-9%) and the lowest for real estate and other debt (less than 2%). On average, less than five percent of asset values are imputed in a given year (lower panel of Table 1). However, for the assessment of total net worth, the number of cases subject to imputation cumulates across wealth components: For about one quarter of cases, at least one wealth component that is part of total net worth is imputed. For about four percent of cases, three or more wealth components are imputed. Finally, in assessments of longitudinal changes (e.g., between two survey waves), the number of cases affected by imputation cumulates across years (see “Overall” column): only 62% of cases require no imputation of any wealth component in either year.

The described random hot-deck imputation was a state-of-the art method in the 1980s. In contrast, modern approaches incorporate covariates to increase the precision of the imputations, e.g., in a regression-switching framework, a technique that would have been all but impossible to implement back then given the limited computing power. The quality of imputed data is known to vary across different imputation approaches (Frick, Grabka, & Sierminska, 2007); the hot-deck imputation approach currently applied by PSID might be particularly prone to inflate estimates of wealth fluctuation, calling for the type of methodological assessment provided here.

3 Analytic Approach, Measures, Methods

3.1 Analytic Approach

We assess the relationship between large wealth fluctuation and potential sources of measurement error, including the number of imputed wealth components. However, we also investigate the extent to which actual changes in households' socio-economic circumstances predict large wealth fluctuation. It is necessary to pursue both aims at the same time. By jointly estimating the conditional role of imputation as a potential source of measurement error on the one hand and substantively meaningful changes on the other, we take into account that the two might be interrelated. For example, item nonresponse might be correlated with turbulences in a household's socio-economic conditions if a respondent is less likely to recall or disclose asset information if he recently lost his job and now consumes out of his family's assets.

It is important to note that our analyses cannot provide a strict comparative adjudication between the total "signal" and "noise" underlying large wealth fluctuation. Although our analyses include another potential source of measurement error, an indicator noting whether there was a change in respondent between waves, we cannot claim to exhaustively capture all possible "noise," nor, for that matter, all possible "signals." Instead, we reveal *some* of the predictors of large wealth fluctuation that likely indicate measurement error to motivate further improvements in data collection and processing. At the same time, we reveal substantively meaningful sources of changes in household wealth, which might – especially if they account for a significant share of large wealth fluctuation – caution against the premature conclusion that large wave-to-wave fluctuation in hard-to-measure economic variables is inherently problematic.

3.2 Sample and Measures

For this methodological project, we use PSID's imputed net worth variables that cumulate all measured asset and debt components (see Table 1) to examine net worth fluctuation between the 2005 and 2007 waves. We selected these two waves to circumvent strong period effects in subsequent waves brought about by the Great Recession in the form of substantial shocks to the wealth holdings of many American families (see Pfeffer, Danziger, & Schoeni, 2013). Our main analytic sample comprises 7,051 households with the same household head at both time points.[2]

Our outcome measures are based on the following six different specifications of wealth changes:

(a) absolute gains and absolute losses in net worth between 2005 and 2007, i.e., W_{2007}-W_{2005} ("absolute gain/loss");

(b) gains and losses in net worth between 2005 and 2007 relative to 2005 net worth among those with positive net worth in both years, i.e., $(W_{2007}$-$W_{2005})/W_{2005}$ ("relative gain/loss (to net worth)"); and

(c) gains and losses in net worth between 2005 and 2007 relative to 2005 household income among those with positive net worth in both years, i.e., $(W_{2007}$-$W_{2005})/I_{2005}$ ("relative gain/loss (to income)").

Though each of these measures has its advantages and disadvantages,[3] as we will show, they yield similar overall conclusions about the determinants of wealth fluctuation.

Determining the degree of wealth fluctuation that is large enough to raise suspicion about its sources is ultimately based on a subjective decision about what constitutes "too" extreme of a change. In this contribution, we define extreme gains and extreme losses as cases within the top five percent of the overall distribution of wealth gains and losses, respectively. Results based on just the top 2.5% yield similar results and are available upon request.

[2] Drawing the analytic sample based on household heads observed in both waves is one common and necessary strategy to identify households across waves. It does, of course, condition on an important aspect of demographic changes in household structure (namely, the dissolution or formation of a household with a new household head) and, as such, provides a conservative estimate of the role of demographic changes in accounting for large wealth fluctuations.

[3] For example, households with greater wealth should be more likely to experience large absolute changes (e.g., losing more than $200,000) whereas households with lower wealth should be more likely to experience large changes relative to their baseline net worth (e.g., double their wealth by moving from $100 to $200 net worth). Additionally, the measure of change relative to baseline household income is also intended to address these distributional concerns (e.g., a wealth gain of $10,000 for a household with an income of $50,000 is treated the same as a wealth gain of $50,000 for a household with an income of $250,000).

Predictors of large wealth fluctuation (i.e., independent variables), include

(a) indicators of measurement characteristics, including the *number of imputed wealth components across both waves* (see Table 1, bottom panel) and whether there was a *change in respondent*,[4]

(b) an encompassing list of demographic characteristics (*age*, *sex*, and *race of household head* and *baseline wealth*) and changes in socio-economic circumstances between 2005 and 2007, including changes in *household composition*, *asset portfolios ("active savings")*, *labor market participation*, and *health conditions* (see Appendix A for a detailed list).

3.3 Methods

To analyze the determinants of large wealth fluctuation, we estimate logistic regression models for each of our six outcome variables (large gains and losses as absolute, relative to net worth, and relative to income changes). All of our analyses are weighted using the 2005 PSID family weight. All regression coefficient estimates are displayed as average marginal effects in Appendix A. For ease of presentation and interpretation, we display a selection of the main estimates in the form of predicted probabilities – more specifically, as discrete changes based on average marginal effects. We also briefly discuss model fit based on a pseudo-R2 for logistic regressions following McKelvey and Zavoina (1975), a measure that has been shown to best approximate the "percent explained variance" interpretation commonly used in OLS regressions (Hagle & Mitchell, 1992; Windmeijer, 1995).[5] All estimates are produced using the `margin` and `spost` commands in Stata 14 (Long & Freese, 2014).

[4] The PSID does not necessarily interview the same respondent in both years, even in households with no composition change since the prior wave. For instance, a husband might be the respondent in one year whereas his wife might be the respondent in another year.

[5] This interpretation requires us to assume a latent trait underlying our outcome variables (Long & Freese, 2014). Such an assumption seems justified in this application because we are more interested in evaluating the latent trait of "wealth fluctuation" than in evaluating the observed trait of specifically falling into the outlying gains/losses of the wealth change distribution. The fit statistics reported here are based on unweighted regressions.

Table 2 Distribution of Wealth Fluctuation

	Absolute Change	Relative Change (to baseline wealth)	Relative Change (to baseline income)
Percentile 1	-1,026,278	-0.97	-30.96
Percentile 5	-218,700	-0.80	-5.50
Median	7,800	0.20	0.33
Percentile 95	525,000	7.00	10.17
Percentile 99	1,885,300	53.55	47.26
N	7,051	5,329	5,323

4 Results

4.1 Distribution of Wealth Fluctuation

Table 2 displays the distributions of our main measures of between-wave wealth fluctuation: absolute change as well as relative change among those with positive net worth in both years (relative to baseline wealth and relative to baseline income). The median wealth change is $7,800 in absolute terms, 20% relative to net worth, and 33% relative to income. Inflation accounts for at least some of the increase in the first two measures; however, we do not adjust for inflation because we are more interested in the accuracy of respondents' reports than relating wealth to changing macro-economic conditions. The typical degree of wealth fluctuation reported here indicates that wealth tended to increase leading up to the crash, a finding consistent with prior research based on the same data (Pfeffer et al., 2013).

Our main interest here is in the tails of the distribution of wealth fluctuation. As shown in Table 2, the largest five percent of wealth losses and gains, which we designate as large fluctuation for the purpose of this contribution, are losses of $218,700 or more and gains of $525,000 or more, respectively. Large fluctuation relative to net worth includes losses of 80% or more and gains by a factor of 7 or more. Large wealth fluctuation relative to income includes cases experiencing a loss of wealth that is at least 5.5 times as high as their baseline income or a gain of wealth that is at least 10.2 times as high as their baseline income.

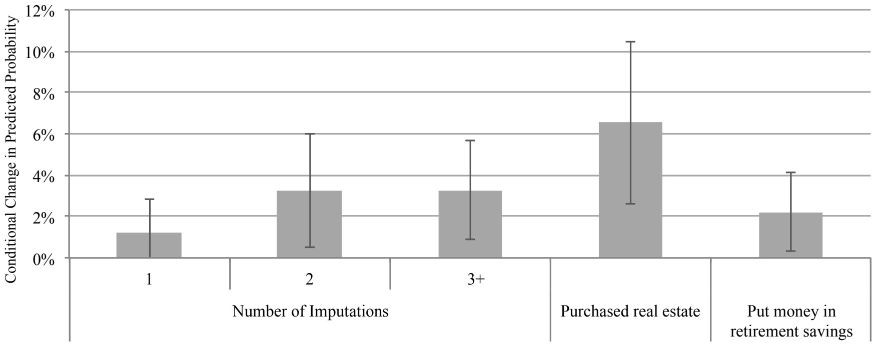
4.2 Predictors of Large Wealth Fluctuation

Table A.1 reports regression estimates for the prediction of large gains and large losses across the different specifications of wealth change (six separate regressions) and reveals the main categories of variables that independently and consistently predict large wealth changes (besides the expected association between baseline level of wealth and wealth changes; see footnote 4): (1) the number of imputations as an indicator of potential measurement error, (2) changes in asset portfolios, and (3) changes in household composition (though we remind the reader that our sample necessarily conditions on some fundamental changes in household composition; see footnote 3). Here, we report some of the main results in graphical form to facilitate interpretation. Specifically, we illustrate those predictors that are generally the largest and most consistent predictors of wealth gains or losses (see Figure 1a for predictors of large absolute gains and Figure 1b for predictors of large absolute losses). Although we only display associations with *absolute* gains and losses, graphs showing (in many cases even larger) associations with relative wealth changes are available upon request.

Figure 1a shows that, conditional on all other observed factors, the imputation of one wealth component (in either year) is associated with an increase in the probability of observing a large absolute wealth gain by 1.2 percentage points, though not statistically significant (p>.05). The imputation of two or more wealth components is associated with a statistically significant increase in the probability of observing large wealth gains by about 3 percentage points (p<.05). That is, we are 3.3 percentage points more likely to observe a large wealth gain for households with at least two imputed wealth components compared to otherwise similar households for whom we observe all wealth components. Because the baseline probability of experiencing a large wealth gain, as defined in this study, is 5 percentage points, an increase of 3 percentage points is substantial. We return to a substantive interpretation of these associations below.

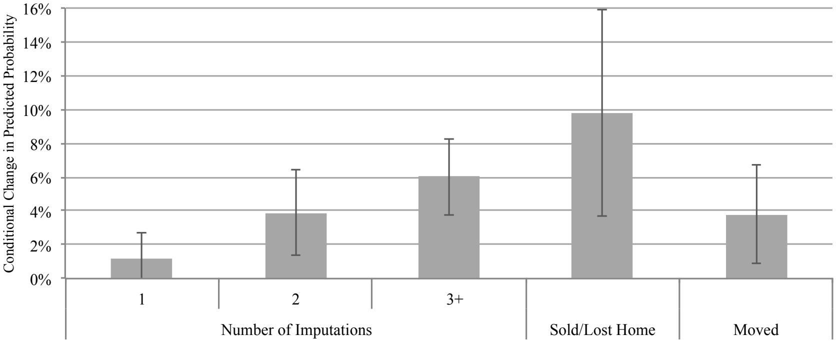
Figure 1a also displays results for two examples of substantively meaningful predictors of large wealth gains: purchasing real estate and saving for retirement. Specifically, everything else equal, the probability of observing a large absolute wealth gain is 6.6 percentage points greater for households that purchased real estate (other than their main residence) and 2.2 percentage points greater for those who put money into retirement savings (private annuities and Individual Retirement Accounts). We then observe that some substantive indicators – such as the purchase of real estate – are more predictive of large wealth gains than the imputation indicator chosen here.

Figure 1b displays the independent predictors of large absolute wealth *losses* and reveals quite similar conclusions. Specifically, everything else equal, the probability of observing large wealth losses is 3.9 percentage points and 6 percentage points greater among those with two imputed wealth components and those with three or more imputed wealth components, respectively, compared to those with no wealth imputations. Furthermore, large wealth losses are also associated with substantively meaningful changes in household characteristics, including the transition from home ownership to non-ownership (with an increase in probability of 9.8 per-



Note: Discrete change (based on logistic regression reported in A.1); 95% confidence intervals

Figure 1a Predictors of Large Positive Wealth Fluctuation (Absolute Gains)



Note: Discrete change (based on logistic regression reported in A.1); 95% confidence intervals

Figure 1b Predictors of Large Negative Wealth Fluctuation (Absolute Losses)

centage points) and, separately, the household moving to a different residence (an increase of 3.8 percentage points).

We judge all conditional associations shown here to be of considerable size. But how do we interpret them in substantive terms? We designated as "substantively meaningful predictors" the various aspects of active savings that are independently associated with large wealth fluctuation, including the purchase of real estate, putting money into retirement savings, and selling or losing a home (see Table A.1 for others, such as the purchase of stocks or home improvements). We believe that these indicators are likely to reflect true fluctuation in households' economic profiles: Some households experience both large wealth shifts and shifts in their wealth portfolio and investment behavior together. However, we do not claim that these factors exert a causal effect; in fact, for many of these factors, it is unclear whether they should be thought of as determinants of a large wealth change (e.g., selling a house in a bad market might trigger a substantial loss of net worth) or a consequence (e.g., the involuntary loss of a house, such as through foreclosure, might be caused by preceding socio-economic troubles and asset losses). Either way, we believe that changes in active savings and several other household characteristics listed in Table A.1 are sources of meaningful wealth fluctuation.

In contrast, we believe that the independent association between wealth fluctuation and the presence of imputations suggest that the imputation algorithm currently applied might be a source of measurement error underlying large wealth changes.[6] Having described the nature of the hot-deck imputation algorithm above, this interpretation seems quite probable to us. Of course, theoretically, the imputation indicator might also be a proxy for selective nonresponse. That is, even with the ample list of observable control variables included here, it is possible that reports on wealth components might not be missing at random (MAR). However, the structure of selective nonresponse would have to be quite peculiar to produce the patterns observed here: similarity in the associations between the imputation indicator and large wealth gains and wealth losses as well as the monotonic increase in the probability of large fluctuation across the number of imputed components.

4.3 Accounting for Large Wealth Fluctuation

In a final step, we evaluate whether the observed household characteristics and potential measurement artifacts studied here account for an appreciable share of the variability in wealth fluctuation. This assessment is based on the estimated pseudo R2 reported in the bottom panel of Table A.1. Across all outcomes, our full models account for a substantial share of the variability in wealth fluctuation and, for half of the models, the majority of the variability (row 1). Indicators of demographic and changes in socio-economic characteristics alone (row 2) explain between one quarter and one half of the variance in wealth gains (38% of absolute gains, 48 % of gains relative to wealth, and 27% of gains relative to income) and up to four fifths of the variance in wealth losses (80% of absolute losses, 31% of losses relative to wealth, and 50% of losses relative of income). As a single predictor, the number of wealth components imputed (row 3) explains up to 11% of the variance whereas a change in respondent explains far less (row 4). However, conditional on the predictors of meaningful wealth fluctuation (row 1), the contribution of measurement error indicators is quite modest (compare rows 1 and 2): The additional variance

[6] We also note that our other tested indicator of measurement error, a change in respondent, is a less consistent predictor of wealth fluctuation. Specifically, a change in respondent independently predicts extreme changes in relative gains but not other specifications of change.

explained by taking into account indicators of potential measurement error is less than 5% for all models and far less in most (about 1%).

5 Conclusion

We have studied between-wave changes in family net worth as an increasingly important indicator of economic wellbeing that is also particularly hard to measure. Using PSID data from 2005 and 2007, we sought to differentiate between substantively meaningful predictors of wealth fluctuation (specifically, changing socioeconomic and demographic conditions of households) and potential measurement error arising from wealth imputations and a change in respondent.

Deciding what degree of wealth change is large enough to qualify as suspicious is arbitrary; here, we focused on the five percent of households that experienced the largest absolute and relative gains and the five percent that experienced the largest losses. Using this definition, we were able to account for between 31% and 80% of large wealth losses (depending on whether measuring absolute or relative losses) and between 29% and 52% of large gains based only on households' demographic and socio-economic characteristics and changes therein. In other words, the mere fact that a household's wealth in one wave is radically different from its wealth in the prior wave should not automatically trigger concerns about the presence of measurement error. Instead, the best explanations for such extreme fluctuation (other than the household's baseline level of wealth) are changes in asset portfolios. For example, a change in home ownership is highly predictive of experiencing large wealth fluctuation as are other asset portfolio changes, such as the purchase of real estate or investments in businesses.

However, we have also shown in detail that, whereas the imputation strategy currently implemented by PSID contributes only a small additional portion to the overall explained variance in wealth fluctuation, having more imputed wealth components is clearly and independently associated with large wealth fluctuation. This finding suggests that the random hot-deck imputations that were the state-of-the-art approach when the PSID began collecting wealth data in the 1980s could be updated to accommodate covariates, including information from prior and subsequent waves (Moldoff et al., 2013; Westermeier & Grabka, 2015). In particular, including the changes in life circumstances identified here (e.g., changes in home ownership and active savings behaviors) appears to be a promising next step in improving the wealth data provided by PSID and perhaps other surveys.

Generally speaking, multivariate multiple imputation methods have been demonstrated to be superior to univariate single imputation methods. For example, in an evaluation of methods for imputing bracketed survey data on household wealth in the Health and Retirement Study, Heeringa, Little, and Raghunathan,

(2002) found that a Bayesian approach to multiple imputation was more effective than complete-case analysis, mean or median substitution, and multiple imputation based on a univariate hot deck (see Heeringa, 1999 for earlier simulation work demonstrating the utility of the method). More recent research directly addresses the effectiveness of incorporating longitudinal information in the imputation of panel data, considering the effects of imputation on both cross-sectional accuracy (e.g., trends, distributions, and measures of inequality) and longitudinal accuracy (e.g., distributional accuracy of wealth mobility). Although Frick & Grabka (2007) found that imputations incorporating longitudinal information were superior to those that did not, Kennickell (2011) found no meaningful differences between different methods and Westermeier & Grabka (2015) found that no single method was best for all scenarios. To that end, future methodological work should explore the effectiveness of a variety of these latest imputation techniques given the particulars of PSID. In the meantime, analysts are able to utilize imputation flags provided by the PSID to re-impute wealth information themselves and, in the process, ensure that their imputation models mimic their specific analytic models (Allison, 2002; Rubin, 1987).

References

Allison, P. D. (2002). *Missing data* (Vol. 136). Thousand Oaks, CA: Sage Publications, Inc.

Bosworth, B. P. & Smart, R. (2009). *Evaluating micro-survey estimates of wealth and saving* (Working Paper 2009-4). Center for Retirement Research at Boston College, Boston.

Bucks, B., & Pence, K. (2015). Wealth, pensions, debt, and savings: Considerations for a panel survey. *Journal of Economic and Social Measurement, 40,* 151-175.

Conley, D. (2001). Capital for college: Parental assets and postsecondary schooling. *Sociology of Education, 74*(1), 59-72.

Deaton, A. (1991). Saving and liquidity constraints. *Econometrica,59,* 1221-48.

Duncan, G. J. (1984). *Years of poverty, years of plenty: The changing economic fortunes of American workers and families.* Ann Arbor, MI: Institute for Social Research, University of Michigan.

Frick, J. R., & Grabka, M. M. (2007). *Item non-response and imputation of annual labor income in panel surveys from a cross-national perspective* (Discussion Paper 763). DIW Berlin, Berlin.

Frick, J. R., Grabka, M. M., & Sierminska, E. M. (2007). *Representative wealth data for Germany from the German SOEP: The impact of methodological decisions around imputation and the choice of the aggregation unit* (Discussion Paper 672). DIW Berlin, Berlin.

Groves, R. M. (2004). *Survey errors and survey costs.* Hoboken, NJ: Wiley-Interscience.

Hagle, T. M., & Mitchell II, G. E. (1992). Goodness-of-fit measures for probit and logit. *American Journal of Political Science, 36*, 762-784.

Heeringa, S. G. (1999). *Multivariate imputation and estimation for coarsened survey data on income and wealth* (Doctoral dissertation). University of Michigan, Ann Arbor, MI.

Heeringa, S. G., Little, R. J. A., & Raghunathan, T. E. (2002). Multivariate imputation of coarsened survey data on household wealth. In R. M. Groves, D. A. Dillman, J. L. Eltinge, & R. J. A. Little (Eds.), *Survey nonresponse* (pp. 357-371). New York: John Wiley & Sons.

Hill, D. H. (2006). Wealth dynamics: Reducing noise in panel data. *Journal of Applied Econometrics, 21*, 845-860.

Juster, F. T., & Smith, J. P. (1997). Improving the quality of economic data. Lessons from HRS and AHEAD. *Journal of the American Statistical Association, 92*, 1268-1278.

Kennickell, A. B. (2011). Look again: Editing and imputation of SCF panel data. In *JSM Proceedings*, Survey Research Methods Section. Alexandria, VA: American Statistical Association.

Long, S., & Freese, J. (2014). *Regression models for categorical dependent variables Using Stata*. College Station: Stata Press.

McKelvey, R. D., & Zavoina, W. (1975). A statistical model for the analysis of ordinal level dependent variables. *Journal of Mathematical Sociology, 4,* 103-120.

Moldoff, M., Meijer, E., Chien, S., Hayden, O., Hurd, M., Main, R., Miu, A., Rohwedder, S., & St. Clair, P. (2013). *RAND HRS income and wealth imputation, Version M.* Rand Center for the Study of Aging, Santa Monica, CA.

Panel Study of Income Dynamics (2013). PSID main interview user manual: Release 2013. Ann Arbor, MI: University of Michigan.

Panel Study of Income Dynamics (2015). Public use dataset. Produced and distributed by the Survey Research Center, Institute for Social Research, University of Michigan, Ann Arbor, MI.

Piketty, T. (2014). *Capital in the twenty-first century.* Cambridge, MA: Belknap Press.

Pfeffer, F. T., & Schoeni, R. F. (eds.) 2016. Wealth inequality: Economic and social dimensions. *Russell Sage Journal of the Social Sciences, 6(2).*

Pfeffer, F. T., Danziger, S. H., & Schoeni, R. F. (2013). Wealth disparities before and after the great recession. *Annals of the American Academy of Political and Social Science, 650*(1), 98-123.

Rubin, D. (1987). *Multiple imputation for nonresponse in surveys*. New York: Wiley.

Venti, S. F. (2011). Economic measurement in the Health and Retirement Study. *Forum for Health Economics & Policy, 14* (3), Article 2.

Westermeier, C., & Grabka, M. M. (2015). *Longitudinal wealth data and multiple imputation: An evaluation study* (SOEPpapers on Multidisciplinary Panel Data Research 790). DIW Berlin, Berlin.

Windmeijer, F. A.G. (1995). Goodness-of-fit measures in binary choice models. *Econometric*

APPENDIX A

Predictors of Large Wealth Fluctuation

Table A.1 Logistic Regressions

Average marginal effects (standard errors in parentheses)

	Gains (top 5%)			Losses (top 5%)		
	Absolute	Relative (to net worth)	Relative (to income)	Absolute	Relative (to net worth)	Relative (to income)
Wealth Imputation						
Number of asset components imputed (reference = none)						
One	0.012 (0.008)	0.020 * (0.008)	0.005 (0.009)	0.012 (0.008)	0.001 (0.008)	0.013 (0.010)
Two	0.033 * (0.014)	0.044 ** (0.014)	0.04 * (0.016)	0.039 ** (0.013)	0.009 (0.014)	0.045 ** (0.016)
Three or more	0.033 ** (0.012)	0.078 *** (0.019)	0.067 *** (0.017)	0.060 *** (0.012)	0.009 (0.013)	0.071 *** (0.015)
Change in Respondent						
Whether respondent changed between 2005 and 2007	0.022 (0.014)	0.026 * (0.010)	0.045 *** (0.013)	-0.009 (0.017)	0.012 (0.015)	-0.012 (0.020)
Demographics						
Age of HH head	0.000 (0.000)	0.000 (0.000)	0.001 *** (0.000)	0.000 (0.000)	-0.001 * (0.000)	0.000 (0.000)
Sex of HH head (reference = female)	0.024 * (0.011)	0.005 (0.008)	-0.005 (0.011)	0.009 (0.010)	-0.017 (0.009)	-0.031 ** (0.010)

	Gains (top 5%)			Losses (top 5%)		
	Absolute	Relative (to net worth)	Relative (to income)	Absolute	Relative (to net worth)	Relative (to income)
Race of HH head (reference = white)						
Black	-0.003 (0.012)	0.003 (0.008)	0.022 (0.016)	0.024 (0.018)	0.015 (0.011)	0.030 (0.019)
Other	-0.013 (0.012)	0.016 (0.010)	0.015 (0.016)	-0.007 (0.012)	0.033 * (0.016)	0.005 (0.017)
Baseline wealth quintile						
1st	reference	reference	reference	reference	reference	reference
2nd	0.004 (0.008)	-0.436 *** (0.053)	-0.044 (0.038)		0.016 (0.012)	
3rd	0.002 (0.006)	-0.593 *** (0.055)	-0.069 (0.037)		0.018 (0.013)	0.008 * (0.004)
4th	0.019 * (0.008)	-0.621 *** (0.056)	-0.045 (0.037)	0.011 *** (0.003)	0.025 (0.014)	0.025 *** (0.007)
5th	0.081 *** (0.011)	-0.627 *** (0.056)	-0.022 (0.038)	0.216 *** (0.020)	0.046 ** (0.016)	0.145 *** (0.016)
Change in household composition						
Any change in household structure	0.002 (0.008)	0.007 (0.007)	-0.019 (0.011)	0.000 (0.009)	0.012 (0.008)	-0.013 (0.011)
Change in sum of assets and/or debts due to mover-in or mover-out (reference = else)						
Net gain	0.050 (0.026)	0.021 (0.018)	0.061 (0.032)			
Net loss				0.021 (0.018)	0.040 * (0.017)	0.016 (0.024)

	Gains (top 5%)			Losses (top 5%)		
	Absolute	Relative (to net worth)	Relative (to income)	Absolute	Relative (to net worth)	Relative (to income)
Whether household member entered college	0.027 *	0.009	0.011	-0.003	-0.007	-0.014
	(0.011)	(0.014)	(0.016)	(0.014)	(0.014)	(0.018)
Whether family moved	0.000	-0.002	0.012	0.030 **	0.033 **	0.037 **
	(0.012)	(0.007)	(0.013)	(0.010)	(0.011)	(0.013)
Changes in asset portfolio						
Whether change in home ownership status (reference = else)						
Now owns home	0.023	0.056 ***	0.057 ***			
	(0.018)	(0.008)	(0.016)			
No longer owns home				0.065 ***	0.094 ***	0.064 ***
				(0.016)	(0.013)	(0.018)
Whether sold home used as main dwelling	-0.034 *	0.021	-0.027	-0.040 **	-0.028	-0.051 *
	(0.017)	(0.011)	(0.021)	(0.015)	(0.016)	(0.022)
Whether purchased real estate other than main home	0.045 ***	0.037 *	0.023	0.009	0.008	-0.009
	(0.010)	(0.017)	(0.015)	(0.011)	(0.017)	(0.016)
Whether sold real estate other than main home	-0.004	-0.009	0.000	0.01	0.016	0.024
	(0.015)	(0.027)	(0.019)	(0.015)	(0.027)	(0.022)
Whether made home additions or improvements	0.026 ***	0.015	0.019	-0.013	-0.057 **	-0.015
	(0.007)	(0.011)	(0.011)	(0.008)	(0.019)	(0.011)
Whether purchased non-IRA stock	0.021 *	-0.003	0.007	-0.014	-0.006	-0.018
	(0.009)	(0.012)	(0.013)	(0.009)	(0.014)	(0.011)
Whether sold non-IRA stock	0.009	0.014	-0.002	0.008	-0.015	0.003
	(0.009)	(0.014)	(0.015)	(0.010)	(0.018)	(0.012)
Whether put money into private annuities or IRAs	0.019 **	0.027 **	0.010	-0.019 **	-0.059 **	-0.022 *
	(0.007)	(0.009)	(0.010)	(0.007)	(0.018)	(0.010)

	Gains (top 5%)			Losses (top 5%)		
	Absolute	Relative (to net worth)	Relative (to income)	Absolute	Relative (to net worth)	Relative (to income)
Whether cashed in any part of pension, private annuity, or IRA	0.017 (0.010)	-0.001 (0.015)	0.010 (0.012)	-0.016 (0.012)	-0.017 (0.017)	-0.025 (0.014)
Whether invested in business or farm	0.020 * (0.010)	0.029 ** (0.010)	0.023 (0.014)	0.010 (0.010)	-0.003 (0.015)	0.013 (0.013)
Whether sold business or farm	-0.020 (0.022)	0.008 (0.026)	-0.016 (0.030)	0.013 (0.025)	0.009 (0.035)	-0.013 (0.031)
Whether received gift or inheritance >=$10k in last two years	-0.002 (0.012)	0.040 ** (0.014)	0.014 (0.020)	0.010 (0.012)	-0.067 ** (0.025)	0.012 (0.014)
Whether received large settlement or inheritance in last year	0.009 (0.012)	0.013 (0.014)	0.010 (0.018)	-0.011 (0.012)	-0.005 (0.016)	-0.005 (0.014)
Changes in labor market participation						
Change in head employment status (reference = else)						
Employed to unemployed				0.017 (0.014)	0.012 (0.014)	0.015 (0.016)
Unemployed to employed	0.006 (0.016)	-0.016 (0.016)	0.026 (0.019)			
Change in spouse employment status (reference = else)						
Employed to unemployed				0.016 (0.013)	0.000 (0.014)	0.019 (0.016)
Unemployed to employed	-0.028 (0.022)	0.018 (0.013)	-0.019 (0.024)			

	Gains (top 5%)			Losses (top 5%)		
	Absolute	Relative (to net worth)	Relative (to income)	Absolute	Relative (to net worth)	Relative (to income)
Whether change in head occupation	0.010 (0.007)	0.000 (0.006)	0.016 (0.009)	-0.002 (0.007)	-0.004 (0.007)	-0.009 (0.008)
Whether change in spouse occupation	0.000 (0.007)	-0.001 (0.007)	-0.009 (0.010)	-0.013 * (0.007)	-0.013 (0.008)	-0.020 * (0.009)
Change in head retirement status (reference = else)						
Retired to not retired	0.018 (0.024)	-0.070 * (0.030)	0.022 (0.022)			
Not retired to retired				-0.008 (0.015)	-0.009 (0.022)	0.007 (0.017)
Change in spouse retirement status (reference = no else)						
Retired to not retired	-0.054 (0.032)	-0.007 (0.035)	-0.024 (0.027)			
Not retired to retired				-0.025 (0.018)	0.038 (0.022)	-0.021 (0.023)
Changes in Health Status						
Change in head health status (reference = else)						
Worse				0.005 (0.008)	-0.003 (0.009)	0.006 (0.009)
Better	0.001 (0.008)	-0.007 (0.008)	0.000 (0.010)			
Change in spouse health status (reference = else)						
Worse				0.009 (0.009)	-0.005 (0.012)	-0.018 (0.013)
Better	-0.001 (0.010)	-0.009 (0.009)	-0.017 (0.014)			

	Gains (top 5%)			Losses (top 5%)		
	Absolute	Relative (to net worth)	Relative (to income)	Absolute	Relative (to net worth)	Relative (to income)
Change in head perceived health status (reference = else)				-0.016	0.005	-0.004
Worse				(0.009)	(0.010)	(0.010)
Better	-0.016 (0.011)	-0.004 (0.008)	-0.030 (0.017)			
Change in spouse perceived health status (reference = else)						
Worse				-0.026 * (0.012)	-0.020 (0.015)	-0.015 (0.017)
Better	-0.001 (0.013)	-0.015 (0.011)	0.001 (0.018)			
Change in health condition limiting work for head (reference = else)						
Worse				0.015 (0.011)	0.005 (0.012)	0.012 (0.013)
Better	-0.037 (0.019)	0.022 (0.014)	-0.015 (0.018)			
Change in health condition limiting work for spouse (reference = else)						
Worse				0.016 (0.015)	-0.037 (0.019)	0.019 (0.018)
Better	-0.007 (0.020)	-0.034 (0.022)	-0.013 (0.023)			

	Gains (top 5%)			Losses (top 5%)		
	Absolute	Relative (to net worth)	Relative (to income)	Absolute	Relative (to net worth)	Relative (to income)
Pseudo R2 (McKelvey & Zavoina)						
(1) All predictors included	0.397	0.520	0.293	0.804	0.309	0.516
(2) Only socio-econ. & demogr. changes	0.382	0.475	0.267	0.803	0.306	0.501
(3) Only imputation indicator	0.040	0.017	0.073	0.076	0.003	0.109
(4) Only change in respondent	0.005	0.005	0.017	0.000	0.003	0.001
N	6,594	4,994	4,988	6,594	4,994	4,988

Note: Statistical significance levels at * p<.05, ** p<.01, and *** p<.001 based on two-tailed tests. For the prediction of large absolute and relative (to income) losses, some bottom wealth quintiles had to be merged due to perfect prediction. The specification of predictors capturing directional change (e.g. health conditions worsening vs. improving) is targeted to the outcome measure (e.g. for wealth losses: health conditions worsening vs. not; for wealth gains: health conditions improving vs. not).

Information for Authors

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

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

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

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

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

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

Entire Book:

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

Journal Article (with DOI):

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

Journal Article (without DOI):

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

Chapter in an Edited Book:

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

Internet Source (without DOI):

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

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

gesis
Leibniz Institute for the Social Sciences

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