mda

methods, data, analyses

Journal for Quantitative Methods and Survey Methodology

Volume 13, 2019 | 2

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.mda.gesis.org

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

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

Print: Bonifatius Druck GmbH Paderborn, Germany

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

Content

RESEARCH REPORTS

RESEARCH NOTES

FIELD REPORTS

Editorial

As a survey methodologist, I always feel that there are not enough journals focusing on the methodological and statistical aspects associated with survey research. The journal *methods, data, analyses (mda)* is one of the few existing exceptions. In addition, *mda* is an open-access journal, which allows results to be shared broadly. Therefore, I am delighted to introduce my first issue as the editor-in-chief of *mda*.

First and foremost, I would like to thank all the associate editors for their crucial support of the journal, and especially Sabine Häder (GESIS – Leibniz Institute for the Social Sciences) for her continuous commitment as managing editor. Moreover, I sincerely thank the previous editor-in-chief, Annelies G. Blom (University of Mannheim), for her great work during the past years to professionalise and internationalise *mda* and to increase its scientific impact.

The introduction of the professional journal management tool "Open Journal System (OJS)" facilitates the management of submissions and makes it easier for authors to track the evolution of their submissions. In addition, the editorial board was extended to efficiently deal with the review process. Furthermore, *mda* established an online first publication section, which allows a timely publication after the acceptance of the manuscripts. This contributes to a fast sharing of knowledge, which is particularly important nowadays since the world is changing so quickly. I am also delighted to announce that *mda* was included in the Emerging Sources Citation Index (ESCI) and that the journal applied for its inclusion in the Social Science Citation Index (SSCI).

In 2018, *mda* published two issues with a total of 11 manuscripts. The acceptance rate was 50%. The number of downloads from our *mda* website, i.e., single articles or whole issues, was 14,657 and the average number of citations per paper was 1.7 (Google Scholar).

In 2019, we expect that the journal will continue improving. The first issue of 2019 (published in January) included seven outstanding research reports from a variety of international authors. This second issue (July 2019) includes four research reports, two research notes, and one field report dealing with issues related to both measurement errors and sampling. A new extension of the editorial board was implemented to integrate experts from emerging research fields, such as the use of sensor data in surveys. In addition, we will keep working on the internationalisation of *mda* and on guaranteeing high quality and timely publications.

However, the success of *mda* also depends on the authors that consider the journal as a good outlet for their research, the research community and readers that use the knowledge gained from the publications, and the anonymous reviewers that provide careful feedback on the manuscripts to help improve them. Therefore, I would like to sincerely thank all authors, readers, and reviewers of *mda* and encourage them to get even more involved in the journal.

Melanie Revilla

methods, data, analyses | Vol. 13(2), 2019, pp. 199-222 DOI: 10.12758/mda.2018.03

Chasing Hard-to-Get Cases in Panel Surveys: Is it Worth it?

Nicole Watson & Mark Wooden

Melbourne Institute of Applied Economic and Social Research, University of Melbourne

Abstract

In many population surveys, fieldwork effort tends to be disproportionately concentrated on a relatively small proportion of hard-to-get cases. This article examines whether this effort is justified within a panel survey setting. It considers three questions: (i) are hard-to-get cases that are interviewed different from other interviewed cases? (ii) do cases that require a lot of effort in one survey wave require a lot of effort in all waves? and (iii) can easy-to-get cases be re-weighted to eliminate biases arising from not interviewing hard-to-get cases? Using data from a large nationally representative household panel survey, we find that hard-to-get cases are distinctly different from easy-to-get cases, suggesting that failure to obtain interviews with them would likely introduce biases into the sample. Further, being hard-to-get is mostly not a persistent state, meaning these high cost cases are not high cost every year. Simulations confirm that removing hard-to-get cases introduces biases, and these biases lead to an understatement of the extent of change experienced by the population. However, we also find that under one of five fieldwork curtailment strategies considered, the bias in population estimates that would arise if the hard-to-get cases were not pursued can be corrected by applying weights. Nevertheless, this conclusion only applies to the curtailment strategy involving the smallest decline in sample size. Biases associated with curtailment strategies involving larger sample size reductions, and hence greatest cost savings, are not so easily corrected.

Keywords: HILDA Survey, sample representativeness, longitudinal surveys, fieldwork curtailment strategies, fieldwork efficiency

The return to additional survey fieldwork effort, as measured by additional survey respondents, invariably declines with the rate of response. Obtaining very high response rates to population surveys thus typically requires concentrating fieldwork effort, especially towards the end of the fieldwork period, on a relatively small proportion of cases who are hard-to-get. But is the extra effort and cost spent on achieving high response rates justified?

Most previous research on the fieldwork effort involved in following up hard-to-get cases has been undertaken within the context of cross-sectional surveys (e.g., Billiet et al., 2007; Fitzgerald & Fuller, 1982; Hall et al., 2013; Heerwegh et al., 2007; Lin & Schaeffer, 1995; Lynn et al., 2002; Stoop, 2005). In these studies, the analysis usually revolves around a comparison of the characteristics of two groups of respondents: those defined as 'hard-to-get' and the remainder. This, of course, ignores the group that objectively are the hardest to get – the non-responders – reflecting the fact that often little is known about non-respondents in cross-section surveys. Of course, some limited information about non-respondents can be garnered from frame characteristics (Etter & Perneger, 1997), though the range of variables available is usually quite limited, or from non-response follow-up studies that investigate the reasons for non-response (Stoop, 2005, pp. 146-156), though these studies also suffer from non-response issues. In contrast, a large amount is usually known about panel survey respondents who subsequently do not respond in a later wave. As a result, considerable research has been undertaken into the causes and consequences of panel attrition (for example, Behr et al., 2005; Lepkowski & Couper, 2002; Lugtig et al., 2014; Watson & Wooden, 2009). Panel surveys also provide a rich setting for examining the effectiveness of fieldwork effort in following up hard-to-get cases each wave. Previous studies on this issue that have involved panel survey data, however, have mostly focused on identifying hard-to-get cases in just one survey wave (e.g., Haring et al., 2009; Larroque et al., 1999; Ullman & Newcomb, 1998). This is surprising given the ramifications for pursuing

Acknowledgments

The research reported on in this paper was, in part, supported by an Australian Research Council Discovery Grant (#DP1095497). This paper uses unit record data from the Household, Income and Labour Dynamics in Australia (HILDA) Survey. The HILDA Project was initiated and is funded by the Australian Government Department of Social Services (DSS) and is managed by the Melbourne Institute of Applied Economic and Social Research (Melbourne Institute). The findings and views reported in this paper, however, are those of the authors and should not be attributed to either DSS or the Melbourne Institute.

Direct correspondence to

Nicole Watson, Melbourne Institute of Applied Economic and Social Research,
Level 5, 111 Barry Street, University of Melbourne, Victoria 3010, Australia
E-mail: n.watson@unimelb.edu.au

or not pursuing cases extend well beyond a single wave. In this article, we examine whether the fieldwork effort devoted to obtaining hard-to-get interviews across six annual survey waves is justified.

Another feature of previous research using either cross-sectional or panel data is the wide variation across studies in how a ‘hard-to-get’ case is defined. The most common types of definitions employed include any case that: requires a large number of number of visits or calls (Cottler et al., 1987; Hall et al., 2013; Heerwegh et al., 2007; Kennickell, 2000; Lin & Schaeffer, 1995; Lynn et al., 2002; Yan et al., 2004); has refused earlier in the fieldwork period (Billiet et al., 2007; Hall et al., 2013; Fitzgerald & Fuller, 1982; Lin & Schaeffer, 1995; Lynn et al., 2002; Woodruff et al., 2000; Yan et al., 2004); or was interviewed late in the fieldwork period (Etter & Perneger, 1997; Haring et al., 2009; Kennickell, 2000; Lahaut et al., 2003; Larroque et al., 1999; Studer et al., 2013; Ullman & Newcomb, 1998; Yan et al., 2004).

Most studies find that hard-to-get cases are different from easy-to-get cases. These differences extend from socio-demographic variables such as age (Cottler et al., 1987; Hall et al., 2013; Kennickell, 2000; Larroque et al., 1999), sex (Cottler et al., 1987), race (Cottler et al., 1987; Hall et al., 2013), and education (Cottler et al., 1987; Etter & Perneger, 1997; Kennickell, 2000; Larroque et al., 1999), to more substantive variables such as employment (Hall et al., 2013), occupation (Larroque et al., 1999), income (Etter & Perneger, 1997; Kennickell, 2000), wealth (Kennickell, 2000), smoking (Woodruff et al., 2000), substance use (Studer et al., 2013), and physical health (Etter & Perneger, 1997). Obtaining interviews with these hard-to-get cases is expected to reduce biases in survey estimates. How important this reduction in bias is, however, depends on how similar the interviewed hard-to-get cases are to the non-respondents.

For longitudinal surveys, decisions about how much effort to devote to pursuing hard-to-get cases should be influenced, at least in part, by expectations about the likelihood of retaining such sample members in subsequent waves. Being a hard-to-get respondent in one wave, for example, has been found to be predictive of attrition in the next (Haring et al., 2009; Watson & Wooden, 2009). More generally, does the extra effort (and cost) required to interview the hard-to-get cases fall persistently on the same cases from wave to wave? As far as we are aware, this is an issue not considered in any previous research.

Further, and perhaps most importantly, relatively few studies have tested in a simulation setting whether re-weighting the easy-to-get cases can reduce the potential biases introduced from not pursuing interviews with the hard-to-get cases. And those studies that have been conducted (e.g., Billiet et al., 2007; Hall et al., 2013) have used cross-sectional data.

This article uses the Household, Income and Labour Dynamics in Australia (HILDA) Survey, a household-based panel study, to examine three related questions.

1. Are hard-to-get cases that are ultimately interviewed different from other interviewed cases?
2. Do cases that require a lot of effort in one survey wave require a lot of effort in all waves?
3. Can easy-to-get cases be re-weighted to eliminate biases potentially arising from not interviewing hard-to-get cases?

We build on previous research in a number of ways. First, we define hard-to-get cases in five different ways and so can assess how sensitive conclusions are to the choice of measure. Second, we analyze the extent to which being hard-to-get is a state that persists over time. Third, we examine whether the biases that may result if fieldwork is curtailed over an extended period (six annual survey waves) can be eliminated by re-weighting the remaining (i.e., easy-to-get) cases.

Data

The HILDA Survey is a panel that began in 2001 with a three-stage stratified clustered nationally representative sample of households (Watson & Wooden, 2012). There were 19,914 people living in the 7682 responding households in wave 1. These people are followed over time and the sample is extended to include all people living with these original sample members at the time of the subsequent interviews. Interviews are conducted annually with all sample members aged 15 years or older. The vast majority (over 90 percent) of these interviews are undertaken face-to-face, with the remainder by telephone.

The initial responding sample was achieved from a total of 11,693 households identified as in-scope, giving a wave 1 household-level response rate of 66 percent (AAPOR RR1). Annual re-interview rates of individuals are high, rising from 87 percent in wave 2 to over 94 percent by wave 5, and remaining above that level in all subsequent waves.

Within each wave of fieldwork there are three distinct phases, with each successive phase increasingly focusing on sample members that are hardest to locate, contact and interview. The initial fieldwork phase is concentrated in August to September. Non-responding and partially responding households are reviewed and reissued for follow-up fieldwork in the next phase (October to December). The third fieldwork phase (in January to February) is used to contact households that were difficult to trace or where it is believed further contact attempts may be successful.

Methods

Defining Hard-to-Get Cases

We examine a range of different definitions of 'hard-to-get', based on the length of time since commencement of fieldwork, whether an initial refusal was received, and the number of calls made. The most natural delineations of time for the HILDA Survey are the fieldwork stages described earlier, with the survey manager deciding at the end of the first and second fieldwork stages who among the non-respondents should be re-approached. The two time-based definitions of hard-to-get cases used here are:

- Definition A: The individual was interviewed during a follow-up stage of fieldwork.
- Definition B: The interview was completed after the New Year.
- Alternatively, the survey manager may choose not to re-issue to field anyone who initially refused. This suggests a third definition:
- Definition C: The individual initially refused before being interviewed.

Finally, we create binary variables based on the number of calls exceeding some threshold. A call is counted if it was a face-to-face visit or if it was a telephone call that resulted in an appointment, an interview, or other information to finalize the outcome of an individual. From 2009 (wave 9), a change from pen-and-paper interviewing to computer-assisted personal interviewing facilitated the collection of detailed call records. Using these records, we can determine the number of calls made to the household before a particular individual is interviewed. The distribution of these calls, based on data pooled from waves 9 to 14, is shown in Figure 1. For this analysis, we focus on two specific thresholds – 7 or more calls, and 13 or more calls required to obtain an interview. Obviously, a number of different thresholds could have been selected due to the greater granularity of call-based measures compared to those used in the first three definitions. The choice of the particular thresholds used here reflects the operational requirements imposed on the company engaged to undertake the fieldwork for the HILDA Survey. Specifically, an interviewer must make at least six calls to a household in a particular fieldwork period before they can return the household to the office with an inconclusive outcome (such as a non-contact), and then up to a further 6 calls after making contact to interview sample members. This provides two further definitions.

- Definition D: 7 or more calls were made to the household by the time the individual was interviewed.
- Definition E: 13 or more calls were made to the household by the time the individual was interviewed.

Figure 1 Distribution of calls made before interviewing sample member, waves 9 to 14 combined

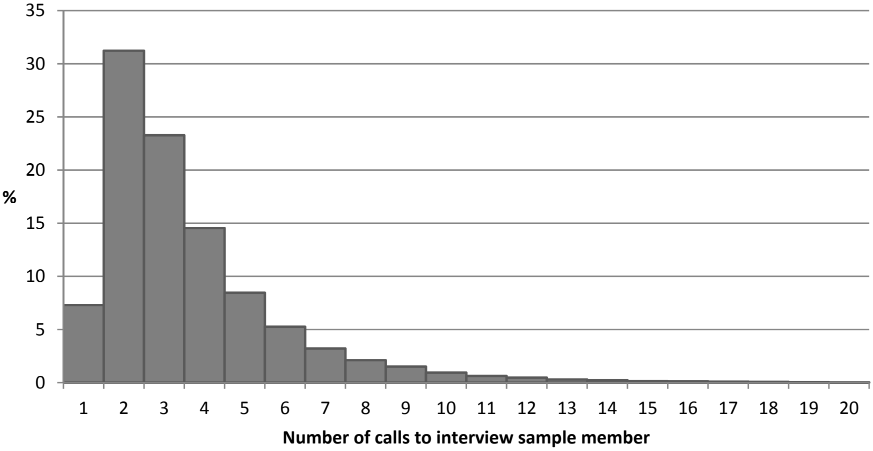


Figure 2 Percentage of Interviewed Cases That Were 'Hard-to-Get', by Wave

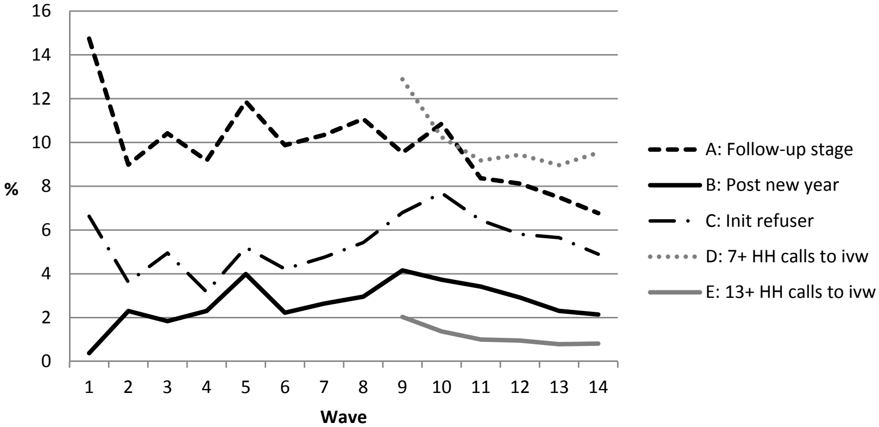


Figure 2 shows the proportion of interviews that were hard-to-get according to each of these five definitions, and how this has varied over time. Approximately 10 percent of interviews required a follow-up period of fieldwork to achieve the interview (definition A); though there is a noticeable decline in this proportion in later waves (waves 11 to 14). Only about half of these follow-up cases were due to an initial refusal (definition C) in the early waves, but this rises to around 70 percent in waves 9 to 14. This shift coincides with a change in fieldwork provider, which occurred after wave 8, suggesting either a change in re-issuing practice or a greater ability on the part of the new provider to convert initial refusals to interviews. The proportion of cases interviewed after the New Year (definition B) each wave is relatively small (2 to 4 percent) and varies somewhat wave to wave. The proportion of cases defined as hard-to-get when using call counts varies substantially depending on the particular call threshold applied. Using a cut-off of 7 or more calls to define a hard-to-get case (definition D) results in 9 to 13 percent of the interviewed cases being classified as hard-to-get. When the higher cut-off of 13 or more calls is used (definition E), the proportion of interviewed cases defined as hard-to-get declines to just 1 to 2 percent.

Assessing the Impact of Pursuing Hard-to-Get Cases

Multinomial logistic models of the three interview outcomes at wave *t* – easy-to-get interview, hard-to-get interview, and not interviewed – are used to assess whether the hard-to-get cases are appreciably different from the easy-to-get cases (research question 1). We include a range of personal and household characteristics, all measured at wave *t*-1, that are often found to be associated with non-response (see Watson & Wooden, 2009). These include: age (in 10-year bands), sex, marital status (6 categories), number of adults living in the household, number of children (aged less than 15) living in the household, education level (6 categories), country / region of birth (3 categories), whether the sample member has a restrictive long-term health condition, area of residence (9 categories), employment status (6 categories), real equivalized (i.e., household size adjusted) gross annual (financial year) household income (with missing values imputed; see Hayes & Watson, 2009), whether an owner-occupier of a home, whether the household moved between waves *t*-1 and *t*, and a set of wave indicators.

Missing data on covariates resulted in the loss of just 520 observations (0.7 percent) from the models employing the first three definitions of hard-to-get, leaving a total of 77,315 person-wave observations. For the last two hard-to-get definitions, a further 54 person-wave observations were dropped due to missing call record information. To allow repeated observations on the same individuals, the multinomial logistic models are fitted as two-level hierarchical models where level 1 is the wave observation and level 2 is the individual. Two random effects, which were allowed to be correlated, were assumed for the different interview outcomes.

To assess whether individuals are hard-to-get repeatedly over time simply because of their particular socio-demographic characteristics (research question 2), we rerun the above set of multinomial logit models and include an indicator variable for whether the individual was hard-to-get in wave *t*-1.

Finally, we test whether excluding the hard-to-get cases materially affects key estimates from the study (research question 3). We examine whether the different sample curtailment strategies are associated with significant differences in selected personal and household characteristics, and assess whether these differences can be eliminated through the application of survey weights constructed for the reduced sample under each of the five curtailment strategies that only contains the easy-to-get cases. We then similarly test for differences in responses to 15 selected estimates of change over time. The weights used relate to a "balanced" panel of respondents from wave 1 to 14 where the hard-to-get cases have been dropped from wave 9 onwards. The balanced panel weights were constructed by adjusting the wave 1 cross-sectional weights for attrition from wave 1 to wave 14 (by multiplying by the inverse of the response propensity that is modelled on a range of wave 1 socio-economic characteristics and some post-wave 1 mobility information where available). The weights are then calibrated to a set of external wave 1 totals. This follows the same methodology employed to construct the regular HILDA Survey weights (Watson, 2012). Standard errors of the difference between the full sample and the truncated sample (i.e., after excluding the hard-to-get cases) for each definition of hard-to-get, were calculated using jackknife estimation with 45 replicates.

To ensure that all definitions are examined across the same timeframe, all analyses that follow are restricted to the outcomes observed in waves 9 to 14.

Results

Are the Hard-to-Get Cases Different From Other Cases?

The coefficients from the estimation of multinomial logit regression models with random effects predicting interview outcomes are shown in Table 1. Separate estimates are provided for each of the five definitions of hard-to-get.

Regardless of the definition used, the hard-to-get group is distinctly different from both the easy-to-get and the non-respondents. Compared to easy-to-get respondents, hard-to-get respondents tend to be younger, single, live in a household with three or more adults, less educated, born in a non-English-speaking country, have higher incomes, not have a restrictive long-term health condition and live in households that have moved. The likelihood of being a hard-to-get case also increases with household income (though at a declining rate) and hours worked. Non-respondents when compared to easy-to-get cases, tend to be relatively young, live in larger households, have not completed high school and likely to live in households that have moved since the previous interview.

Table 1 Multinomial Logit of Interview Outcome, with Random Effects: waves 9 to 14

	A: Follow-up stage		B: Post New Year		C: Initial refuser		D: 7+ calls to interview		E: 13+ calls to interview	
Characteristics from interview in wave *t*-1	Hard	Non-resp	Hard	Non-resp	Hard	Non-resp	Hard	Non-resp	Hard	Non-resp
Age group (base=15-24)										
25-34	-0.056	-0.087	-0.044	-0.096	-0.034	-0.084	0.035	-0.085	0.017	-0.101
35-44	-0.052	-0.003	-0.183	-0.030	-0.050	-0.007	-0.027	-0.017	0.014	-0.009
45-54	-0.117	-0.312**	-0.250*	-0.333**	-0.195*	-0.331**	-0.094	-0.340**	0.094	-0.334**
55-64	-0.279**	-0.809**	-0.487**	-0.830**	-0.494**	-0.841**	-0.430**	-0.899**	-0.373	-0.840**
65+	-0.303**	-0.254#	-0.304#	-0.220	-0.563**	-0.288#	-0.734**	-0.333*	-0.382	-0.243#
Female	0.019	0.025	0.071	0.035	0.071	0.028	-0.043	0.025	-0.070	0.030
Marital status (base=married)										
De facto	0.091	0.142	0.185#	0.145	-0.013	0.130	0.209**	0.152#	0.316*	0.110
Separated	0.092	0.091	0.381*	0.113	0.042	0.062	0.259*	0.152	0.391	0.149
Divorced	-0.142	0.003	0.109	0.040	-0.227*	-0.021	0.168#	0.070	0.511*	0.047
Widowed	-0.160	0.410*	0.059	0.401*	-0.118	0.410*	0.223#	0.522**	-0.210	0.447**
Never married & not living with partner	0.247**	0.347**	0.404**	0.327**	0.042	0.300**	0.361**	0.369**	0.721**	0.298**
Number of adults (base=1 adult)										
2 adults	0.099	0.118	0.106	0.114	0.037	0.111	0.247**	0.161	0.188	0.101
3 adults	0.227**	0.174#	0.285*	0.173#	0.207*	0.160	0.505**	0.242*	0.622**	0.165
4 or more adults	0.150#	0.217#	0.325**	0.243*	0.146#	0.211#	0.716**	0.337**	0.724**	0.213#
Number of children (base=0 children)										
1 child	-0.036	-0.093	-0.060	-0.097	0.023	-0.090	0.024	-0.108	0.203	-0.095
2 children	-0.022	-0.178#	-0.092	-0.197*	0.096	-0.155	-0.023	-0.215*	0.213	-0.193*
3 or more children	-0.016	-0.355**	-0.009	-0.315*	0.041	-0.339*	0.037	-0.382**	0.497**	-0.331*

Table 1 continued

	A: Follow-up stage		B: Post New Year		C: Initial refuser		D: 7+ calls to interview		E: 13+ calls to interview	
Characteristics from interview in wave *t*-1	Hard	Non-resp	Hard	Non-resp	Hard	Non-resp	Hard	Non-resp	Hard	Non-resp
Highest level of education (base=Year 11 or below)										
Year 12	0.103#	-0.134	-0.036	-0.176*	0.025	-0.144	0.078	-0.153#	-0.124	-0.173*
Cert III or IV	0.083	-0.072	0.094	-0.085	0.039	-0.077	0.120*	-0.062	-0.136	-0.096
Diploma	0.025	-0.268*	-0.163	-0.311*	-0.051	-0.291*	0.048	-0.260*	0.079	-0.275*
Graduate	-0.154*	-0.449**	-0.290*	-0.491**	-0.375**	-0.492**	-0.127*	-0.435**	-0.464**	-0.457**
Post graduate	-0.182*	-0.625**	-0.226#	-0.640**	-0.365**	-0.659**	-0.220**	-0.629**	-0.411*	-0.606**
Country of birth (base=Australia)										
Main English-speaking country	-0.034	-0.130	-0.130	-0.152	-0.143	-0.157	-0.017	-0.140	-0.034	-0.143
Not main English-speaking country	0.416**	0.377**	0.432**	0.350**	0.271**	0.338**	0.431**	0.382**	0.537**	0.305**
Long term health condition	-0.114*	-0.010	-0.269**	-0.009	-0.139**	-0.023	-0.127**	-0.015	-0.215#	-0.011
Area of residence (base=Major city: Sydney)										
Major city: Melbourne	-0.164*	-0.187#	-0.257*	-0.170	-0.168*	-0.185#	-0.131*	-0.197#	-0.305*	-0.169#
Major city: Brisbane	-0.086	-0.161	-0.204	-0.160	-0.185#	-0.187	-0.667**	-0.277*	-1.105**	-0.179
Major city: Adelaide	-0.062	-0.082	-0.197	-0.095	-0.186#	-0.099	-0.521**	-0.175	-1.563**	-0.121
Major city: Perth	-0.108	0.072	-0.215	0.071	-0.271*	0.022	-0.296**	0.016	-0.527**	0.058
Major city: other	-0.423**	-0.478**	-0.416**	-0.470**	-0.545**	-0.487**	-0.855**	-0.580**	-1.134**	-0.459**
Inner regional	-0.048	-0.109	-0.326**	-0.137	-0.165*	-0.131	-0.680**	-0.250**	-1.041**	-0.151
Outer regional	0.182*	0.301**	0.202#	0.264*	0.026	0.270*	-0.722**	0.131	-0.727**	0.262*
Remote	0.629**	0.686**	1.004**	0.699**	0.608**	0.694**	-0.205#	0.498**	-0.308	0.536**
Employment status (base=not in labour force)										
Employed, working <=34 hrs	-0.057	-0.127	-0.100	-0.114	-0.006	-0.133	0.100*	-0.125	0.121	-0.128
Employed, working 35-44 hrs	0.087	0.264**	0.143	0.270**	0.208**	0.252**	0.134*	0.273**	0.365*	0.269**

Table 1 continued

	A: Follow-up stage		B: Post New Year		C: Initial refuser		D: 7+ calls to interview		E: 13+ calls to interview	
Characteristics from interview in wave *t*-1	Hard	Non-resp	Hard	Non-resp	Hard	Non-resp	Hard	Non-resp	Hard	Non-resp
Employed, working 45-54 hrs	0.204**	0.321**	0.241*	0.313**	0.308**	0.319**	0.255**	0.342**	0.424*	0.331**
Employed, working 55+ hrs	0.480**	0.609**	0.381**	0.564**	0.542**	0.588**	0.441**	0.587**	0.568**	0.570**
Unemployed	0.018	0.238#	0.066	0.246*	0.143	0.235#	0.152#	0.257*	0.186	0.261*
Equivalized HH income ($/10^5$)	0.435**	0.151	0.402**	0.181	0.382**	0.168	0.443**	0.157	0.317	0.125
Equivalized HH income squared ($/10^{10}$)	-0.049**	-0.003	-0.046	-0.008	-0.038#	-0.005	-0.071**	-0.004	-0.091	-0.002
Owner occupier	0.015	-0.103	-0.014	-0.097	0.033	-0.094	-0.148**	-0.132*	-0.267*	-0.121#
Household moved (btw *t*-1 and *t*)	0.533**	0.419**	0.839**	0.418**	0.605**	0.426**	0.502**	0.398**	0.575**	0.322**
Next wave (base=wave 9)										
Wave 10	0.303**	0.180*	0.065	0.113	0.286**	0.168*	-0.218**	0.058	-0.420**	0.096
Wave 11	0.020	0.170*	0.067	0.168*	0.155*	0.201*	-0.307**	0.104	-0.485**	0.159*
Wave 12	0.053	0.339**	-0.061	0.319**	0.111#	0.358**	-0.198**	0.289**	-0.518**	0.320**
Wave 13	-0.019	0.373**	-0.188*	0.362**	0.124*	0.410**	-0.222**	0.328**	-0.582**	0.367**
Wave 14	-0.115*	0.344**	-0.302**	0.339**	-0.022	0.377**	-0.143**	0.331**	-0.585**	0.354**
Constant	-3.489**	-4.703**	-4.939**	-4.732**	-3.766**	-4.719**	-2.734**	-4.526**	-5.675**	-4.594**
Var (random effects)	2.155**		3.202**		2.411**		1.547**		3.403**	
Cov (random effects for hard and non-resp)	2.728**		3.496**		3.003**		2.181**		3.041**	
Log-likelihood	-29,694		-18,890		-25,832		-31,852		-14,753	
N(person-year observations)	77,315		77,315		77,315		77,261		77,261	

Note: # p<0.10; * p<0.05; ** p<0.01.

How Persistent are Hard-to-Get Cases?

Do cases that require a lot of work in one wave require a lot of work in all waves? Figure 3 shows the proportion of hard-to-get cases at one wave that are interviewed in subsequent waves but were hard-to-get. It shows that the level of reoccurrence is relatively low, with 9 to 24 percent of hard-to-get cases in one wave classified as hard-to-get in the next wave, and the rate of persistence in being classified as hard-to-get declines over time, with 5 to 17 percent classified as hard-to-get four waves later. The large majority (75 to 90 percent, depending on the hard-to-get definition used) of hard-to-get cases are classified as easy-to-get in the next wave.

Does the relatively small amount of persistence observed in the hard-to-get cases remain after controlling for respondent characteristics? To test this, we modify the model presented in Table 1 (which predicts whether a case will be easy-to-get, hard-to-get or a non-respondent) and include an indicator of whether the individual was hard-to-get in the prior wave (when the other characteristics included in the model were measured). Table 2 reports the estimated coefficients and mean predicted probabilities for this variable. We find a strong negative association between being hard-to-get in one wave and being easy-to-get in the next. The predicted probability, holding all else constant, of being an easy-to-get case (using definition A; i.e., whether they require follow-up work or not) for those who were easy-to-get in the previous wave is 89.1 percent. This compares with 80.9 percent of those who were previously hard-to-get. The differences in the predicted probabilities are similar for the other four definitions; i.e., 7.2 percentage points for definition B, 8.4 percentage points for definition C, 8.3 percentage points for definition D, and 6.5 percentage points for definition E.

In summary, the large majority of hard-to-get cases (over 80 percent under all definitions) are easy-to-get come the next survey wave. This is not to say, however, that there is no state persistence; a hard-to-get case is still much more likely (around twice as likely) to be hard-to-get next wave than an otherwise comparable case classified as easy-to-get.

Can the Differences in Hard-to-Get Cases be Corrected by Weighting?

Finally, we examine whether the differences between estimates obtained using only the easy-to-get cases and those obtained using both the easy-to-get and hard-to-get cases can be eliminated by applying weights generated specifically for each truncated sample. We first consider the impact fieldwork effort has on the personal and household characteristics included in Table 1 (with the exception of residential mobility, which is included later in Table 4). Table 3 reports the unweighted and weighted estimates for these variables, measured as of 2014 (i.e., in wave 14).

Figure 3 Average Percentage of Hard-to-Get Interviewed Cases in Future Waves Conditional on Being Hard-to-Get in Wave *t*

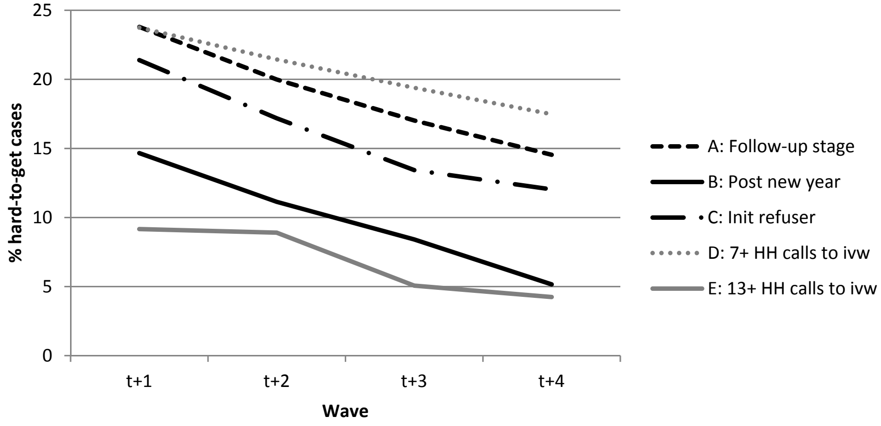


Table 2 Coefficient and Predicted Probabilities for Hard-to-Get in Prior Wave in Multinomial Logit Model of Interview Outcome with Random Effects

	Easy-to-get at *t*			Non-respondent at *t*		
		Mean predicted probability			Mean predicted probability	
Hard-to-get in wave *t*-1	Coeff	Hard at *t*-1	Easy at *t*-1	Coeff	Hard at *t*-1	Easy at *t*-1
Definition A: Follow-up stage	-0.800**	80.9	89.1	0.059	6.9	4.0
Definition B: Post New Year	-1.205**	86.2	93.4	-0.238*	7.9	4.2
Definition C: Initial refuser	-0.936**	82.4	90.8	0.003	7.3	4.1
Definition D: 7+ calls to interview	-0.714**	80.2	88.4	0.181*	6.9	3.8
Definition E: 13+ calls to interview	-1.315**	88.5	95.0	-0.188	9.2	4.2

Note: Models include controls for all the covariates shown in Table 1. # p<0.10; * p<0.05; ** p<0.01.

The weights used to calculate these estimates are for the sample of individuals interviewed in both wave 1 and wave 14, and interviewed in every intervening wave. We describe this as the full balanced panel, though strictly speaking the sample is not completely balanced – respondents that moved abroad and subsequently returned to Australia were also retained. For the curtailed samples, cases that were hard-to-get in any wave between 9 and 14 are dropped from the balanced panel. The full balanced panel from wave 1 to 14 includes 6707 individuals. This declines to 6572 when cases requiring 13 or more calls (definition E) are excluded (a 2 percent reduction), or 6245 cases if the post New Year fieldwork (definition B) is dropped (a 7 percent reduction). Greater reductions in the sample occur when the broader definitions of hard-to-get are used. The balanced panel contains 5661 cases if all initial refusers (definition C) are dropped (a 16 percent reduction), 5225 cases if all follow-up fieldwork (definition A) is abandoned (a 22 percent reduction), or 5046 cases if cases requiring 7 or more calls (definition D) are dropped (a 25 percent reduction).

The unweighted and weighted estimates for the personal and household variables for the full balanced panel are presented in the first two columns of Table 3. The weighted estimates are constructed by weighting the responses provided by both easy- and hard-to-get cases by the wave 1 to 14 balanced panel weight available in the HILDA Survey dataset. The unweighted estimates are similarly restricted to cases that have a positive balanced panel weight to aid comparison of the weighted and unweighted estimates. The following columns in the table provide (for each definition of hard-to-get): i) the difference between the unweighted estimate for the full balanced panel and the unweighted estimate obtained after dropping the relevant hard-to-get cases from waves 9 to 14; and ii) the difference between the weighted estimate for the full balanced panel and the estimate obtained by applying the recalculated balanced panel weight after dropping the relevant hard-to-get cases from waves 9 to 14. The estimates are marked to indicate the p-value for the two-sided z-test for whether this difference is statistically different from zero (# $p<0.10$; * $p<0.05$; ** $p<0.01$).

We find that the definition of hard-to-get that shows the largest number of differences in the unweighted estimates is the curtailment strategy that drops the most cases (definition D which drops cases requiring 7 or more calls) and is least able to be corrected by the weights. The curtailment strategy affecting the personal and household estimates the least is definition E, which drops people requiring 13 or more calls. Further, these estimates are most amenable to correction by the application of weights (while one estimate is not corrected, this is expected by chance alone). Nevertheless, this strategy involves a very small decline in the number of cases followed, and hence the potential for costs savings is commensurately small. Arguably, our results suggest that the best curtailment strategy in terms of maximising sample reduction (and thus saving fieldwork effort) while minimising the effect on estimates is strategy A (not pursuing persons into the follow-up fieldwork phase). However, application of weights is still unable to correct for differences observed on at least three variables (age, country of birth and income).

Next we focus on a subset of variables that relate to change over time, some of which have been much analyzed by users of the HILDA Survey data. The first five estimates relate to changes in the family: the proportion who got married in the last five years; the proportion who separated from a marriage or were widowed in the last 5 years; the proportion that began a de facto relationship in the last year; the proportion who had a new birth in the last year; and the proportion who had a new birth in the last 5 years. There is one measure relating to income: the increase in the 5-year average income between the start and end of the panel (i.e., 2001-05 versus 2010-14). There are four estimates related to employment: whether a new job was started in the last year; whether retired in the last year; for those self-employed in 2009, the proportion that switched to being an employee by 2014; and for those who were employees in 2009, the proportion that transitioned to self-employment by 2014. In terms of health, we include the proportion of people who experienced the onset of a long-term health condition between 2009 and 2014. The final group of four estimates relate to housing: the proportion who moved house in the last year; the proportion who moved house in the last five years; the proportion who transitioned from living in a home that was not owned (i.e., was rented or provided rent-free) to one that was owned between 2009 and 2014; and the proportion who transitioned from living in a home that was owned to one that was not between 2009 and 2014.

The population estimates for the subset of variables are presented in the first column of Table 4. These estimates are constructed by weighting the responses provided by easy- and hard-to-get cases (the full balanced panel) by the wave 1 to 14 balanced panel weight available in the HILDA Survey dataset. Subsequent columns in the table provide (for each definition of hard-to-get): i) the difference between this first population estimate and the one obtained by applying the recalculated balanced panel weight after dropping the relevant hard-to-get cases from waves 9 to 14; and ii) the p-value for the two-sided z-test for whether this difference is statistically equivalent to zero.

We find that the impact of dropping the hard-to-get cases on the selected 15 population estimates is minimal when using the definition involving the loss of fewest cases (definition E); the estimated differences are both very small and a long way from statistically significant. Use of any of the other four definitions, which all involve larger sample losses, results in larger changes in the population estimates. Interestingly, curtailment strategy A, which above we suggested was the best strategy in terms of maximising the reduction in fieldwork effort while having the least impact on the estimates, now appears to be the one that results in the most harm to these estimates of change over time. In general, there is evidence of biases in favour of stability rather than change. For example, the increase in 5-year average income is understated by 6 percent under definition B, 7 percent under definition D, and 10 percent under both definitions A and C. Similarly, under all four of these curtailment strategies the easy-to-get cases are significantly less likely to move house and to separate from marriages.

Table 3 Estimates of Wave 14 (2014) Personal and Household Characteristics and the Effect of Excluding Hard-to-Get Cases

	Estimate based on full balanced panel		Difference in estimate when hard-to-get cases excluded									
			A: Follow-up stage		B: Post New Year		C: Initial refuser		D: 7+ calls to interview		E: 13+ calls to interview	
Characteristic	Unwtd	Wtd	Unwtd[a]	Wtd[b]	Unwtd[a]	Wtd[b]	Unwtd[a]	Wtd[b]	Unwtd[a]	Wtd[b]	Unwtd[a]	Wtd[b]
Age group (base=15-24)												
25-34	9.4	13.9	-0.5**	-0.3**	-0.5**	-0.2#	-0.4*	-0.2*	-0.9**	-0.2	-0.1	0.0
35-44	16.4	18.9	-0.4	-0.4*	-0.2	-0.1	-0.9**	-0.3#	-1.3**	-0.7**	-0.1	-0.1
45-54	23.3	20.6	0.1	0.4	0.2	0.2	-0.2	0.4#	-1.3**	0.2	-0.2#	0.0
55-64	21.6	20.1	0.0	0.2	0.2	0.2	0.3	0.1	-0.1	0.2	0.1#	0.2#
65+	29.4	26.5	0.8*	0.1	0.3#	-0.1	1.1**	0.1	3.6**	0.5#	0.2**	0.0
Female	54.6	51.3	0.0	0.0	0.1	0.0	-0.1	-0.1	0.7**	0.1	0.0	0.0
Marital status (base=married)												
De facto	9.7	10.6	-0.2	-0.2	-0.1	-0.1	-0.1	0.1	-0.9**	-0.7*	0.0	0.0
Separated	3.6	3.4	0.0	0.0	0.0	0.0	-0.1	-0.1	-0.3#	-0.2	0.0	0.0
Divorced	8.5	7.6	0.1	0.0	0.1	0.1	0.0	-0.1	-0.1	-0.3	0.0	0.0
Widowed	7.8	7.1	0.1	-0.2	-0.1	-0.2	0.1	-0.2	0.7**	0.0	0.1**	0.0
Never married & not living with partner	9.6	11.5	-0.2	-0.2	-0.3**	-0.2	0.0	0.0	-0.3	0.1	0.0	0.0
Number of adults (base=1 adult)												
2 adults	54.6	56.5	-0.5	-0.2	0.0	0.2	0.1	0.0	0.7#	-0.1	0.0	-0.1
3 adults	12.6	12.9	0.5*	0.5	0.2*	0.3#	0.2	0.5#	-0.3	0.3	-0.1	-0.1
4 or more adults	10.3	8.9	0.1	0.1	0.2#	0.2	-0.2	-0.1	-0.9**	-0.2	0.0	0.1

Table 3 continued

Characteristic	Estimate based on full balanced panel		Difference in estimate when hard-to-get cases excluded									
			A: Follow-up stage		B: Post New Year		C: Initial refuser		D: 7+ calls to interview		E: 13+ calls to interview	
	Unwtd	Wtd	Unwtd^a	Wtd^b	Unwtd^a	Wtd^b	Unwtd^a	Wtd^b	Unwtd^a	Wtd^b	Unwtd^a	Wtd^b
Number of children (base=0 children)												
1 child	9.8	10.7	-0.2	-0.2	-0.1	-0.1	-0.4*	-0.2	-0.5*	0.0	-0.1	0.0
2 children	10.0	11.3	0.2	0.4	0.1	0.5**	-0.2	0.2	-0.8**	-0.2	0.1	0.1*
3 or more children	4.5	5.2	0.0	0.1	-0.1	-0.1	-0.2	0.0	0.2	0.5**	-0.1	-0.1
Highest level of education (base=Year 11 or below)												
Year 12	10.9	12.4	-0.1	-0.1	-0.1	-0.1	-0.3	-0.1	-0.6*	-0.5#	0.0	0.0
Cert III or IV	22.4	22.9	-0.1	0.0	0.0	0.0	-0.3	-0.2	0.0	0.2	0.0	0.0
Diploma	10.6	10.4	-0.2	-0.1	0.1	0.2	-0.1	0.1	-0.3	0.1	0.0	0.1
Graduate	13.9	14.3	0.0	-0.2	0.1	0.3	0.1	0.0	-0.8*	-0.4	0.0	0.0
Post graduate	12.7	11.7	-0.2	-0.1	-0.3*	-0.4*	0.2	0.1	-0.6*	0.1	0.0	0.0
Country of birth (base=Australia)												
Main English-speaking country	11.1	10.3	-0.1	0.0	0.0	-0.1	0.1	-0.1	0.1	0.0	0.0	0.0
Not main English-speaking country	11.2	14.8	-0.5#	-1.0**	-0.2#	0.0	-0.2	-0.6*	-1.0**	-1.4**	0.0	-0.1
Long term health condition	36.4	35.3	1.1**	0.7#	0.4*	0.2	0.8**	0.3	2.5**	1.1**	0.3**	0.1
Area of residence (base=Major city: Sydney)												
Major city: Melbourne	16.1	17.1	1.0**	0.3#	0.2	0.1	0.5*	0.3	-1.6**	-0.2	-0.2	0.0
Major city: Brisbane	8.0	8.3	-0.2	0.2	-0.1	0.0	-0.2	0.2	0.1	0.0	0.0	0.0
Major city: Adelaide	5.9	5.2	-0.3	0.0	0.0	0.0	-0.1	0.0	0.1	0.0	0.1*	0.0
Major city: Perth	6.8	6.8	-0.2	-0.1	-0.1	-0.2	0.0	-0.2	-0.6**	-0.5*	0.0	0.0

Table 3 continued

	Estimate based on full balanced panel		Difference in estimate when hard-to-get cases excluded									
			A: Follow-up stage		B: Post New Year		C: Initial refuser		D: 7+ calls to interview		E: 13+ calls to interview	
Characteristic	Unwtd	Wtd	Unwtd[a]	Wtd[b]	Unwtd[a]	Wtd[b]	Unwtd[a]	Wtd[b]	Unwtd[a]	Wtd[b]	Unwtd[a]	Wtd[b]
Major city: other	8.7	9.1	0.5*	0.3	0.0	0.1	0.4*	0.2	0.9**	0.4#	0.0	0.1
Inner regional	27.3	23.8	-0.2	-0.2	0.2	0.1	-0.1	-0.2	1.8**	0.6#	0.2**	0.1
Outer regional	11.3	10.0	-0.6#	-0.2	0.0	0.1	-0.3	0.0	1.1**	0.5*	0.1	0.0
Remote	1.4	1.4	-0.2#	-0.1	-0.1	0.0	-0.1	-0.1	0.0	0.0	0.0	0.0
Employment status (base=not in labour force)												
Employed, working <=34 hrs	19.1	18.2	0.4*	0.5#	0.0	0.1	0.1	0.4#	-0.3	0.5	0.0	0.0
Employed, working 35-44 hrs	23.5	25.4	-0.5#	-0.4	-0.2	-0.1	-0.7**	-0.6	-1.8**	-1.0*	0.0	0.0
Employed, working 45-54 hrs	10.2	11.1	-0.6**	-0.4	-0.1	0.0	-0.4**	-0.2	-0.8**	0.2	-0.1#	0.0
Employed, working 55+ hrs	5.6	5.8	-0.4#	-0.2	-0.1	-0.1	-0.2#	-0.1	-0.7**	-0.5#	-0.1	-0.1
Unemployed	1.7	1.8	-0.1	-0.1	0.0	0.1	-0.1#	0.0	-0.3**	-0.2#	0.0	0.0
Equivalized HH income (/10^5)	44,382	45,057	-1601**	-1627**	-791#	-917*	-1183**	-1264*	-2017**	-1337*	-280*	-153
Owner occupier	77.1	74.1	0.2	0.7#	0.4*	0.1	0.3	0.2	1.3**	1.4**	0.1	-0.1

Notes: a – The difference between the unweighted estimate using easy- and hard-to-get cases and the unweighted estimate that would be obtained if no interviews had been completed with the hard-to-get cases. b – The difference between the weighted (population) estimate using easy- and hard-to-get cases and the weighted (population) estimate that would be obtained if no interviews had been completed with the hard-to-get cases. # p<0.10; * p<0.05; ** p<0.01.

Table 4 Population Estimates of Selected Wave 14 (2014) Characteristics Specific to Change and the Effect of Excluding Hard-to-Get Cases

		Difference in population estimate when hard-to-get cases excluded									
		A: Follow-up stage		B: Post New Year		C: Initial refuser		D: 7+ calls to interview		E: 13+ calls to interview	
Population characteristic	Pop. estimate based on full balanced panel	Diff.[a]	p-value	Diff.[a]	p-value	Diff.[a]	p-value	Diff.[a]	p-value	Diff.[a]	p-value
Family											
Got married in last 5 years (%)	6.4	-0.3	0.246	-0.2	0.367	-0.2	0.351	-0.4	0.199	-0.1	0.367
Separated / widowed from marriage in last 5 years (%)	5.5	-0.8	0.000	-0.4	0.011	-0.8	0.000	-0.7	0.003	-0.1	0.427
Began de facto relationship in last year (%)	1.3	-0.2	0.133	-0.1	0.178	-0.1	0.575	-0.1	0.327	0.0	0.778
Had child in last year (%)	2.3	0.0	0.797	-0.1	0.451	0.0	0.876	0.1	0.643	0.0	0.450
Had child in last 5 years (%)	11.8	0.0	0.934	-0.1	0.397	-0.2	0.290	0.1	0.873	-0.1	0.378
Income											
Increase in 5-year average financial year income 2001-05 to 2010-14 ($)[b]	8123	-792	0.013	-510	0.021	-790	0.017	-587	0.048	-226	0.177
Employment											
Started new job in last year (%)[c]	12.4	-0.4	0.306	-0.3	0.231	-0.2	0.514	-0.4	0.516	-0.1	0.348
Retired in last year (%)	1.3	-0.2	0.048	0.0	0.370	-0.1	0.233	-0.2	0.102	0.0	0.324
Transition from self-employed in 2009 to employee in 2014 (%)[d]	30.7	0.8	0.656	-0.4	0.568	0.9	0.441	-1.1	0.548	-0.6	0.298
Transition from employee in 2009 to self-employed in 2014 (%)[e]	3.9	-0.1	0.579	-0.2	0.200	-0.3	0.244	-0.1	0.725	-0.1	0.315

Table 4 continued

Population characteristic	Pop. estimate based on full balanced panel	Difference in population estimate when hard-to-get cases excluded									
		A: Follow-up stage		B: Post New Year		C: Initial refuser		D: 7+ calls to interview		E: 13+ calls to interview	
		Diff.[a]	p-value	Diff.[a]	p-value	Diff.[a]	p-value	Diff.[a]	p-value	Diff.[a]	p-value
Health											
Onset of long term health condition b/w 2009 and 2014 (%)	17.7	0.4	0.417	0.0	0.892	0.2	0.525	0.6	0.202	0.2	0.143
Housing											
Moved b/w 2009 and 2014 (%)	39.7	-1.7	0.000	-0.5	0.026	-1.3	0.000	-1.7	0.001	-0.2	0.091
Moved b/w 2013 and 2014 (%)	12.3	-0.7	0.041	-0.2	0.173	-0.1	0.734	-0.5	0.145	0.0	0.825
Transition from not living in owned home in 2009 to living in owned home in 2014 (%)[f]	28.7	1.7	0.028	0.5	0.251	0.9	0.189	2.6	0.005	-0.1	0.809
Transition from living in owned home in 2009 to not living in owned home in 2014 (%)[g]	8.3	-0.4	0.293	-0.3	0.118	-0.4	0.257	-0.5	0.189	0.1	0.187

Notes: a – The difference between the population estimate using easy- and hard-to-get cases and the population estimate that would be obtained if no interviews had been completed with the hard-to-get cases. b – Gross (i.e., before tax) annual (measured over a financial year; i.e., 1 July to 30 June) real (measured in 2001 prices) personal income. c – Employed persons only. d – Self-employed persons only. e – Employees only. f – Non-home owners only. g – Home owners only.

Discussion

This paper has examined the effect of pursuing hard-to-get cases in a panel setting. We used data from waves 9 to 14 of the HILDA Survey and applied five different definitions of being hard-to-get (based on time in field, whether a refusal was initially obtained, or the number of calls required to achieve the interview). Using different definitions provides a test of the sensitivity of the findings to different possibilities of curtailing the fieldwork effort. Our results suggest three key findings.

First, survey respondents who are hard-to-get, regardless of the definition used, are distinctly different from those who are easy-to-get. This means that in pursuing the hard-to-get cases, we are not simply bringing into the sample more of the same and thus replicating the biases that exist in the sub-sample of early-to-get cases.

Second, being hard-to-get is mostly not a persistent state. The vast majority of sample members who are hard-to-get in one wave (80 to 90 percent) will be easy-to-get in the next wave. This suggests that difficulty obtaining interviews with a case in one wave is largely situational and such cases will not routinely be difficult to interview over a longer time span.

Third, we have uncovered evidence that it is possible to curtail some elements of fieldwork – notably capping the number of call attempts to no more than 12 – without noticeably affecting population estimates. That is, any biases that might arise can be largely rectified through the use of appropriate sample weights. This conclusion, as might be expected, applies to the definition of hard-to-get involving the smallest decline in sample size. When we consider other more significant curtailment strategies involving greater sample losses, and hence greater cost savings, however, the effects on population estimates are more serious. The sample that is lost through these more expansive curtailment strategies tends to be those who have experienced greater change in their lives. Even with the curtailment strategy involving the second smallest decline in the sample size (via dropping the post New Year fieldwork) where the wave-specific estimates can be corrected by weighting, the estimates relating to change over time could not be. Of course, it is not just the number of cases that are dropped that is important, but also what type of cases are dropped. A limitation inherent in examining different definitions of ‘hard-to-get’ is that they will result in different numbers of cases being dropped. However, for the two curtailment strategies that did involve a similar decrease in the number of cases (A and D), we find evidence of different impacts. That is, the curtailment strategy that restricts the number of calls to 6 resulted in substantially more differences in the unweighted and weighted wave-specific estimates but fewer differences in estimates of change over time than the strategy that involved no follow-up fieldwork.

So has devoting effort to chasing hard-to-get cases in the HILDA Survey been worth it? Our answer is a qualified yes. Hard-to-get cases have characteristics that are, on average, quite different from other respondents, suggesting that failure to obtain interviews with them would likely introduce biases into the sample. At the same time, most of these more costly cases are not high cost every year. One qualification is that our simulations suggest that the number of calls to a household could be limited to 12 without significant losses to the sample integrity. This strategy, however, results in a relatively modest reduction in overall sample size (just 2 percent). That said, it is also important to bear in mind that we have only examined the effect of curtailment on a limited set of population estimates; it may be that even very modest curtailment strategies could have significant effects on other estimates.

We expect that these findings are relevant to other longitudinal surveys that employ face-to-face, telephone and possibly even online methodologies. While the definitions of hard-to-get versus easy-to-get may need to change (especially for online surveys), this study has shown that the findings are similar across definitions. Not pursuing the hard-to-get cases could cause biases in estimates that are not able to be eliminated through weighting, and these biases tend to favour stability rather than change over time. We encourage researchers to replicate this analysis with other longitudinal studies. We also encourage use of other definitions of 'hard-to-get', such as the number of calls to first contact and the use of reminder emails or texts (in online surveys).

Finally, we note that we have restricted our attention to potential fieldwork modifications that standardize fieldwork protocols across all cases. An alternative, known as responsive design, is to focus the extended effort only on those cases thought most likely to reduce the bias in key estimates or improve the efficiency of the estimates (Groves & Heeringa 2006; Schouten, Peytchev, & Wagner 2017; Tourangeau et al. 2016). This, however, is far from straightforward in longitudinal surveys or in surveys that cover a wide number of subject domains. Another challenge for all curtailment strategies is that it would require survey funders to shift their focus from response rates as a measure of survey quality to other quality measures (Kreuter 2013).

References

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

Billiet, J., Philippens, M., Fitzgerald, R., & Stoop, I. (2007). Estimation of nonresponse bias in the European Social Survey: Using information from reluctant respondents. *Journal of Official Statistics*, 23(2), 135-162.

Cottler, L. B., Zipp, J. F., Robins, L. N., & Spitznagel, E. L. (1987). Difficult-to-recruit respondents and their effect on prevalence estimates in an epidemiologic survey. *American Journal of Epidemiology*, 125(2), 329-139. doi:10.1093/oxfordjournals.aje.a114534

Etter, J., & Perneger, T. V. (1997). Analysis of non-response bias in a mailed health survey. *Journal of Clinical Epidemiology*, 50(10), 1123-1128. doi:10.1016/S0895-4356(97)00166-2

Fitzgerald, R., & Fuller, L. (1982). I hear you knocking but you can't come in: The effects of reluctant respondents and refusers on sample survey estimates. *Sociological Methods and Research*, 11(1), 3-32. doi:10.1177/0049124182011001001

Groves, R. M., & Heeringa, S. G. (2006). Responsive design for household surveys: Tools for actively controlling survey errors and costs. *Journal of the Royal Statistical Society, Series A*, 169(3), 439-457. doi:10.1111/j.1467-985X.2006.00423.x

Hall, J., Brown, V., Nicolaas, G., & Lynn, P. (2013). Extended field efforts to reduce the risk of non-response bias: Have the effects changed over time? Can weighting achieve the same effects? *Bulletin de Méthodologie Sociologique*, 117(1), 5-25. doi:10.1177/0759106312465545

Haring, R., Alte, D., Völzke, H., Sauer, S., Wallaschofski, H., John, U., & Schmidt, C. O. (2009). Extended recruitment efforts minimize attrition but not necessarily bias. *Journal of Clinical Epidemiology*, 62(3), 252-260. doi:10.1016/j.jclinepi.2008.06.010

Hayes, C., & Watson, N. (2009). HILDA imputation methods. HILDA Project Technical Paper Series 2/09, Melbourne Institute of Applied Economic and Social Research, University of Melbourne. Available at http://melbourneinstitute.unimelb.edu.au/assets/documents/hilda-bibliography/hilda-technical-papers/htec209.pdf

Heerwegh, D., Abts, Koen., & Loosveldt, G. (2007). Minimizing survey refusal and noncontact rates: Do our efforts pay off? *Survey Research Methods*, 1(1), 3-10. doi:10.18148/srm/2007.v1i1.46

Kennickell, A. B. (2000). What do the "late" cases tell us? Evidence from the 1998 Survey of Consumer Finances. Presented at the 1999 International Conference on Survey Nonresponse, Portland, Oregon. Available at www.federalreserve.gov/Pubs/OSS/oss2/papers/icsn99.9.pdf.

Kreuter, F. (2013). Facing the nonresponse challenge. *The Annals of the American Academy of Political and Social Science*, 645(1), 23-35. doi:10.1177/0002716212456815

Lahaut, V. M. H. C. J., Jansen, H. A. M., Van de Mheen, D., Garretsen, H. F. L., Verdurmen, J. E. E., & Van Dijk, A. (2003). Estimating non-response bias in a survey on alcohol consumption: Comparison of response waves. *Alcohol and Alcoholism*, 38(2), 128-134. doi:10.1093/alcalc/agg044

Larroque, B., Kaminski, M., Bouvier-Colle, M., & Hollebecque, V. (1999). Participation in a mail survey: Role of repeated mailings and characteristics of nonrespondents among recent mothers. *Paediatric and Perinatal Epidemiology*, 13(2), 218-233. doi:10.1046/j.1365-3016.1999.00176.x

Lepkowski, J. M., & Couper, M. P. (2002). Nonresponse in the second wave of longitudinal household surveys. In R. M. Groves, D. A. Dillman, J. L. Eltinge and R. J. A. Little (Eds.), *Survey Nonresponse* (pp. 259-272). New York: John Wiley and Sons.

Lin, I., & Schaeffer, N. C. (1995). Using survey participants to estimate the impact of nonparticipation. *Public Opinion Quarterly*, 59(2), 236-258. doi:10.1086/269471

Lugtig, P., Das, M., & Scherpenzeel, A. (2014). Nonresponse and attrition in a probability-based online panel for the general population. In M. Callegaro, R. Baker, J. Bethlehem, A. S. Göritz, J .A. Krosnick and P. J. Lavrakas (Eds.), *Online Panel Research: A Data Quality Perspective* (pp. 135-153). Chichester: Wiley.

Lynn, P., Clarke, P. S., Martin, J., & Sturgis, P. (2002). The effects of extended interviewer efforts on nonresponse bias. In R. M. Groves, D. A. Dillman, J. L. Eltinge, & R. J. A. Little (Eds.), *Survey Nonresponse* (pp. 135-147). New York: Wiley.

Schouten, B., Peytchev, A., & Wagner, J. (2017). *Adaptive Survey Design*. Boca Raton: CRC Press.

Stoop, I. A. L. (2005). *The Hunt for the Last Respondent: Nonresponse in Sample Surveys*. The Hague: Social and Cultural Planning Office of the Netherlands.

Studer, J., Baggio, S., Mohler-Kuo, M., Dermota, P., Gaume, J., Bertholet, N., Daeppen, J., & Gmel, G. (2013). Examining non-response bias in substance use research – are late respondents proxies for non-respondents? *Drug and Alcohol Dependence*, 132(1), 316-323. doi:10.1016/j.drugalcdep.2013.02.029

Tourangeau, R., Brick, J. M., Lohr, S., and Li, J. (2016). Adaptive and responsive survey designs: a review and assessment. *Journal of the Royal Statistical Society, Series A*, 180(1), 203-223. doi:10.1111/rssa.12186

Ullman, J. B., & Newcomb, M. D. (1998). Eager, reluctant, and nonresponders to a mailed longitudinal survey: Attitudinal and substance use characteristics differentiate respondents. *Journal of Applied Social Psychology*, 28(4), 357-375. doi:10.1111/j.1559-1816.1998.tb01710.x

Watson, N. (2012). Longitudinal and cross-sectional weighting methodology for the HILDA Survey. HILDA Project Technical Paper Series 2/12, Melbourne Institute of Applied Economic and Social Research, University of Melbourne. Available at http://melbourneinstitute.unimelb.edu.au/assets/documents/hilda-bibliography/hilda-technical-papers/htec212.pdf.

Watson, N., & Wooden, M. (2009). Identifying factors affecting longitudinal survey response. In P. Lynn (Ed.), *Methodology of Longitudinal Surveys* (pp. 157-181). Chichester: Wiley.

Watson, N., & Wooden, M. (2012). The HILDA Survey: A case study in the design and development of a successful household panel study. *Longitudinal and Life Course Studies,* 3(3), 369-381. doi:10.14301/llcs.v3i3.208

Woodruff, S. I., Conway, T. L., & Edwards, C. C. (2000). Increasing response rates to a smoking survey for U.S. Navy enlisted women. *Evaluation and the Health Professions,* 23(2), 172-181. doi:10.1177/016327870002300203

Yan, T., Tourangeau, R., & Arens, Z. (2004). When less is more: Are reluctant respondents poor reporters? *Proceedings of the Survey Research Methods Section, American Statistical Association*, 4632-4651.

methods, data, analyses | Vol. 13(2), 2019, pp. 223-252 DOI: 10.12758/mda.2018.01

Willingness of Online Panelists to Perform Additional Tasks

Melanie Revilla[1], *Mick P. Couper*[2] & *Carlos Ochoa*[3]

[1] *RECSM-Universitat Pompeu Fabra, Barcelona, Spain*
[2] *University of Michigan, USA*
[3] *Netquest, Barcelona, Spain*

Abstract

People’s willingness to share data with researchers is the fundamental raw material for most social science research. So far, survey researchers have mainly asked respondents to share data in the form of answers to survey questions but there is a growing interest in using alternative sources of data. Less is known about people’s willingness to share these other kinds of data. In this study, we aim to: 1) provide information about the willingness of people to share different types of data; 2) explore the reasons for their acceptance or refusal, and 3) try to determine which variables affect the willingness to perform these additional tasks.

We use data from a survey implemented in 2016 in Spain, in which around 1,400 panelists of the Netquest online access panel were asked about their hypothetical willingness to share different types of data: passive measurement on devices they already use; wearing special devices to passively monitor activity; providing them with measurement devices and then having them self-report the results; providing physical specimens or bodily fluids (e.g. saliva); others. Open questions were used to follow up on the reasons for acceptance or refusal in the case of the use of a tracker.

Our results suggest that the acceptance level is quite low in general, but there are large differences across tasks and respondents. The main reasons justifying both acceptance and refusal are related to privacy, security and trust. Our regression models also suggest that we can identify factors associated with such willingness.

Keywords: online panel; respondent willingness; passive data collection; mobile data collection

The widespread adoption of digital technologies, especially those available on mobile devices, is expanding opportunities for survey researchers to enhance and extend survey measurement, whether through active or passive measurement (see Link et al., 2014). Much of the early research on exploiting these technologies for research has focused on small groups of volunteers. The challenge remains of using these features in the context of large-scale survey data collection. This paper extends that work by exploring stated willingness to provide a variety of types of additional information in the context of an opt-in panel in Spain. We explore willingness across different types of requests that vary in the level of effort required and the degree of intrusiveness, to investigate what additional tasks respondents find more or less acceptable. We also explore reasons for willingness or unwillingness to accept one particular task, installing software to passively track browsing behavior. Finally, we examine the correlates of willingness to accept these additional tasks.

Background

The expansion of the Internet and the development of a range of new active and passive measurement tools, particularly on mobile devices, present a number of potentially exciting opportunities for survey researchers. As the AAPOR (2014) task force noted, “there are a wide array of applications and features available on these devices which can augment and in some cases even replace survey data” (see also Link et al., 2014). The AAPOR report addressed five potential uses of technologies to extend or replace surveys: 1) location or geo-positioning, 2) scanning and QR/barcode readers, 3) visual data capture (photos or video), 4) Bluetooth enabled devices and related technologies, and 5) mobile applications or “apps”. The report calls for “more assessments of auxiliary data collection capabilities,” specifically in terms of “respondent cooperation and compliance, data quality, and potential

Acknowledgments

We thank Gerardo Ortiz for his work on the programming of the survey and Oriol Bosch for his help with coding the open questions. We also thank the Editor, Prof. Edith de Leeuw, and three anonymous reviewers for their very useful comments on a previous draft of this paper.

Direct correspondence to

Melanie Revilla, RECSM-UPF, Edifici Mercè Rodoreda 24.406, Ramón Trías Fargas, 25-27, 08005 Barcelona, Spain
E-mail: melanie.revilla@upf.edu

sources of error" (AAPOR, 2014, p. 9). This paper is focused on the first of these issues.

There are several potential advantages of these new measurement opportunities for supplementing survey data, whether on mobile devices or on PCs. These include 1) reducing respondent burden (i.e., replacing survey questions with passive measurement, or providing easier ways to share information), 2) improving the quality of measurement (i.e., obtaining data that respondents find difficult to report or recall accurately), and 3) measuring new things (i.e., enhancing and extending measurement into new domains).

While a number of papers have argued for the benefits of the enhanced measurement capabilities (see, e.g., Palmer et al., 2013; Raento, Oulasvirta, and Eagle, 2009; Wrzus and Mehl, 2015), much of the research to date has focused on relatively small samples of volunteers. A key challenge for the broader adoption of these new measurement tools in large-scale surveys relates to respondents' willingness to install apps, activate passive tracking, or do the additional tasks researchers ask of them. Further, for those tasks involving ongoing actions beyond the initial consent and installation, continued compliance with the request (or adherence to the protocol) is an additional concern. The mode in which the request is embedded may also be important: Burton (2016) reports a 34 percentage point lower consent rate to administrative record linkages among those surveyed online than those interviewed face-to-face in the *Understanding Society* Innovation Panel.

Several studies have begun to explore these issues and test the feasibility of such additional tasks in the context of ongoing surveys. Most of these studies focus on a single task or technology. For example, some have explored willingness to permit GPS capture. Armoogum and colleagues (2013) asked respondents in the 2007-8 French National Travel Survey (a face-to-face survey) about their willingness to use a GPS device. About one-third (30%) said yes without conditions, while 5% agreed as long as they could turn it off (the rest said no). Biler, Senk, and Winkerova (2013) asked respondents in a face-to-face survey in the Czech Republic about willingness to participate in a travel survey using GPS tracking. Only 8% said they were willing, with 25% uncertain, and 57% not willing. Joh (2017) reports on a pilot study using mail survey recruitment to a travel survey using a GPS-based smartphone app. Of those invited, 5.9% responded to the baseline survey. Of those who reported having a qualifying smartphone, 31.7% downloaded the app and provided some data (representing 1.3% of the original sample).

Turning to online recruitment, Toepoel and Lugtig (2014) asked Dutch panelists for the one-time capture of GPS coordinates: 26% of smartphone participants and 24% of PC participants agreed to such capture. In an online panel study of college students in the U.S., Crawford et al. (2013) reported that 58% said yes to a hypothetical question about GPS capture. In a subsequent wave, between 20% and 33% of survey respondents (depending on the consent condition) provided usable

GPS data. The LISS Mobile Mobility Panel in the Netherlands recruited panelists with smartphones to provide GPS data. Of those who completed the invitation survey (75% of invitees), 37% were willing to participate and 30% (81% of those willing) downloaded the app, activated Wi-Fi and GPS, and provided data for at least one day (Scherpenzeel, 2017). Invitations were restricted to those willing to use their smartphones for research; Antoun, Couper, and Conrad (2017) found that about 41% of LISS panelists were willing to use their smartphones for research.

Other studies have examined the installation of a research app. McGeeney and Weisel (2015) report on a study in the Pew American Trends Panel. Panelists who used an eligible smartphone were randomized to a browser- or app-based version of an experience sampling survey, in which they were asked to complete a short survey twice a day for 7 days. For those in the app group, 76% agreed, and 80% of those installed the app (i.e., 61% of those invited). Completion rates for the 14 surveys were significantly lower for the app group than the browser group. Johnson, Kelley, and Stevens (2012) explored a modular survey design that required installation of an app. Of the panelists who met the eligibility criteria (including use of a smartphone), 43% expressed willingness to do the modular survey and were sent a link to download the app. Of these, 37% (or 16% of qualified panelists) successfully downloaded the app, and 33% (14% of qualified panelists) completed one or more surveys.

A few studies have attempted the collection of passive tracking (e.g., browser log) data. For example, de Reuver and Bouwman (2015) tried to recruit participants from a Dutch online access panel. An initial random sample of the panel did not yield sufficient panelists willing to install the tracker. They then targeted panelists who had previously agreed to the collection of log data. Among these, 31% expressed willingness to allow capture of log data, 22% installed the app, and 14% participated for the full four weeks of the study. The primary reason for non-participation provided was related to privacy (16% of those who provided a reason), followed by a variety of situational factors (holidays, illness, etc.; 15%). Reasons for dropping out during the study were primarily related to technical issues such as battery drainage and reduced performance of the phone. Van Duivenvoorde and Dillon (2015) asked eligible respondents in an opt-in panel in the U.S. to participate in a follow-up study which required them to install passive tracking software. Among respondents who completed the baseline survey (32% of those invited), 3.6% expressed willingness and 2.1% installed the software. Kissau and Fischer (2016) similarly invited members of a Swiss panel to install tracking software. Of those invited, 23% of the main and 8% of the boost sample respectively expressed interest in the study, while 10% of the main and 3% of the boost sample installed the tracking app. In a similar study in Spain using the same tracking app (Wakoopa; see https://wakoopa.com/) Revilla, Ochoa, and Loewe (2017) reported that between 30% and 50% of loyal panelists who were invited agreed to install the tracker.

The capture of accelerometry data (often using stand-alone devices) is more common in large face-to-face surveys, with wide variation in agreement and compliance rates. For instance, Lauderdale et al. (2014) report an initial agreement rate of 80.3% for a sleep actigraphy study, with 88.4% of those who consented (69.8% of those invited) providing usable data. Roth and Mindell (2008) reported a similar consent rate of 80.3%, with 47.7% of those consenting (38.3% of the initial sample) providing usable data (for other examples, see Hassani et al., 2014; Menai et al., 2017; Gilbert et al., 2017). Howie and Straker (2016) conducted a review of trials involving accelerometer use among children, and reported compliance rates ranging from 2% to 60%.

We know of only two studies that has attempted accelerometry measurement in an online study. The FLASHE study in the U.S. (see https://cancercontrol.cancer.gov/flashe) recruited dyads of caregivers and their 12-17 year-old children participants from a commercial online access panel. Of those invited, 39% consented and were enrolled in the study. Of those who consented and were randomly assigned to the survey and accelerometer study, 59% completed the study (23% of invitees). In contrast, for those assigned to the survey-only group, 86% completed the study. Scherpenzeel (2017) reports a 57% willingness and 51% adherence rate to an accelerometry study in the Dutch LISS panel.

Even more intrusive biomarker measures are often used in face-to-face surveys (e.g., McFall, Conolly, and Burton, 2014; Sakshaug, Couper, and Ofstedal, 2010), or as a follow up to telephone surveys (e.g., Boyle et al., 2010; Gautier et al. 2016), but few studies have tested biomarker measures in the context of Internet surveys. For instance, while biomonitors are increasingly being used to study alcohol consumption among volunteers (see, e.g., Greenfield, Bond, and Kerr, 2015), we know of no studies that have tested this on general population samples, especially those with online participants. In one exception, Avendano, Scherpenzeel, and Machenbach (2011) undertook a small pilot in the LISS panel. Panelists were recruited for home cholesterol measurement, involving a finger-prick and blood spot measurement using a device designed for self-administration. Of the 200 panelists invited, 38 (19% of invitees) returned a blood sample, 31 of whom (16% of invitees) provided valid data. Another subsample was asked to chew on a cotton swab and return the saliva sample for cortisol measurement. Of the 200 invited, 30 (15% of invitees) completed the task.

Gatny, Couper, and Axinn (2013) tested the collection of saliva in an ongoing Internet diary study of young women. Saliva kits were mailed to 150 respondents who reported the end of a romantic relationship and were eligible to participate in the collection; 65% mailed back a saliva sample. Similarly, Etter and Bullen (2011) recruited 196 users of electronic cigarettes online, and mailed them a saliva kit: 16% mailed back a saliva sample.

Two other recent studies are relevant. Jäckle and colleagues (2017) invited panelists in the *Understanding Society* Innovation Panel to download an app to scan receipts and report their spending over 4 weeks; 16.5% of respondents downloaded the app and completed the registration survey, and 12.8% used it at least once. Similarly, Angrisani, Kapteyn, and Samek (2017) invited panelists in the Understanding America Study (UAS) to sign up to a customized financial aggregator website and provide financial information. Of those invited, 65% consented; 68% of those who consented (32% of those invited) signed up, and 38% of those (12% of those invited) linked one or more financial institutions.

This brief review of selected studies shows a wide range of stated or actual willingness across a variety of tasks and settings. Many of the studies report rates of compliance without looking at reasons behind the decision or examining differences between those who are or are not willing (some exceptions are reviewed below). Further, all these studies examine only a single request for additional data or technology use.

In one of the few studies to both explore reasons for unwillingness and examine socio-demographic correlates, Pinter (2015) asked members of an access panel in Hungary who used smartphones if they were willing to install a research app. In response to the initial request, 42% said they were unwilling, with a further 23% uncertain (the remaining 35% were willing). Those who were uncertain or unwilling to install the research app were asked their reasons for not being willing, in a series of closed questions. Major reasons proffered by this group included (multiple mentions possible): not enough free time (61%); not enough information to decide (53%); concerns about extra costs of using an app (45%); would participate in some research activities but not others (44%); and concerns about battery use (43%). After additional persuasion aimed at these concerns, 57% eventually agreed to install the app. Pinter also found that behavioral variables (frequency of smartphone use, number of apps, use of GPS, etc.) were moderately but significantly correlated with willingness. In addition, significant but weak correlations of willingness with other socio-economic variables (including political orientation, age, labor force status, income, and frequency of socialization) were found.

Armoogum et al. (2013) examined demographic correlates of willingness to use a GPS device. They found that younger persons, males, those in smaller households, with higher income, with a computer in the household, and with more cars were more willing to participate. Biler et al. (2013) reported that those who used a shopping or travel discount card were more willing to agree to GPS tracking, as were those who used navigation features on their smartphone, those who use social networks, younger persons, and those in larger households.

In a study among Netquest panelists in seven countries, Revilla and colleagues (2016) elicited panelists' willingness to do three additional tasks: 1) share GPS location, 2) install an app, and 3) take a photo. Focusing on the data from Spain, 36% of smartphone owners who responded to the survey said they were definitely willing to install an app, while a further 27% said they were probably willing, and only 15% said definitely no. They found consistent significant negative effects of age on tolerance for additional tasks across countries, but not for other variables (gender, education, and household size) in multivariable models.

Keusch et al. (2017) used a vignette approach to vary features of the request to install a tracking app in a study among opt-in panel members in Germany. Overall they found that 64.5% would *not* be willing (0-5 on an 11-point scale), and 34.9% would *definitely not* be willing (0 on the scale) to install a tracking app. Factors that affected willingness included the sponsor of the study, the length of time that the tracker would be used, the size of the incentives, and the ability to turn off the tracker.

Wenz, Jäckle, and Couper (2017) measured willingness to perform a variety of task in the *Understanding Society* Innovation Panel in the U.K. They found that willingness varied by task (e.g., 59.3% accelerometry capture, 36.7% GPS capture, and 25.5% tracking app). They found lower willingness for men, those with lower education, and those with higher security concerns. Jäckle and colleagues (2017) explored demographic and behavioral correlates of participation in a spending app study. They found that frequency of Internet and mobile device use, along with general cooperativeness with research, were predictors of participation in the app study.

This review of the emerging literature illustrates some of the challenges of exploiting the technical capabilities of modern technology to enhance and extend measurement. The studies reveal considerable variation in willingness to use new technologies for research. But because each study looks at only a single technology or data collection activity, it is hard to determine if this variation is due to the type of request or other features of the design (such as the mode in which the request is made, or the sample on which the study is based). Further, there are inconsistent findings with regard to socio-demographic correlates of willingness, which again may vary by task. Relatively little attention has focused on behavioral and attitudinal correlates of willingness.

This paper adds to this literature by examining panel members' stated willingness to perform a variety of additional tasks. This allows us to explore variation both between tasks and between respondents. We expect the key factors affecting the decision to accept a task include privacy concerns and the effort (or burden) required. Because of the concern for privacy, we expect tasks in which respondents have control of the information provided to have higher levels of acceptance than tasks in which the information is provided automatically to the panel company. However, because of the level of effort required, we expect passive measurement to have higher levels of acceptance than active measurement.

Our focus here is on *stated* willingness, rather than *actual* compliance with the request. We expect that actual compliance rates will be lower than those based on expressed willingness. However, research has shown that stated willingness is a useful measure in its own right, especially if the goal is to examine reasons for and covariates of (un)willingness (see, e.g., Couper and Singer, 2013; Couper et al., 2008, 2010).

Methodology

Data

We use data on the self-reported willingness to complete a variety of different tasks in a web survey. The data was collected from the 15th of September to the 3rd of October 2016 using the Netquest opt-in panel in Spain (www.netquest.com).

Since 2014, Netquest has invited selected panelists to install a tracker (or "meter") on the devices that they are using to go online (PCs, tablets or smartphones), and share (passively) with Netquest the information registered by this tracker (URLs of the web pages visited, time of the visits, ad exposure, and app use in the case of mobile devices) (see Revilla, Ochoa, and Loewe, 2016). In this paper, we only considered panelists who had not yet been invited to install the tracker. In addition, because the survey was also used for other experiments[1], it focused on panelists who have Internet access through both a PC and smartphone. Panel profile information was used to send the invitation to panelists meeting this criterion, and filter questions were used to verify such access. Cross quotas for age and gender were used to ensure that the distribution of these variables in the sample was similar to that observed in the full panel.

The survey contained a maximum of 69 questions. Respondents were able to proceed without providing an answer to the questions; however, they were not able to go back to a previous question. In addition to the questions on willingness to participate in different tasks, the survey included questions on trust and personality traits, as well as socio-demographic questions, and questions about the survey experience/context. The full survey (in Spanish) can be found at the following link: http://ww2.netquest.com/respondent/glinn/mobile2016.

In this paper, we focus on two sets of 10 questions on the willingness to participate in different tasks. These questions were asked in different ways to different respondents, who were randomly assigned to answer through grids or item-by-item questions with vertical or horizontal scales; and on a PC or smartphone. Random-

[1] The other experiments compare the answers for PC and smartphone respondents to rank order questions, grids versus item-by-item questions, and agree-disagree versus item-specific formats.

ization was done independently for each experiment, which allows us to test for confounding. A series of Kolmogorov-Smirnov tests for equality of distributions across the different groups showed significant differences in only a very few cases (2 out of 60 showed significant effects for device; 0 out of 40 for grid versus item-by-item; and 1 out of 40 for scale direction). Thus, we see no evidence of confounding, and ignore the other experiments in our subsequent analyses. We describe the 20 questions in more detail below.

From the 5,907 panelists invited to the survey, 3,051 started it (51.7%) but only 1,623 (53.2% of those who started) answered the first main survey question (following the screener questions). The rest were screened out for a variety of reasons (e.g., used a different device, did not have Internet access through both PC and smartphone, quotas full). Another 132 respondents were excluded later because they switched device during the survey or did not pass some basic quality checks (e.g., the answers to the gender and/or age questions differed from the profile information). Finally, 15 participants dropped out after the first four questions. Thus, a total of 1,476 respondents (48.4% of those who started; 90.9% of those who answered the first main survey question) finished the survey using the required device type; these are the focus of our analyses[2].

Data Preparation and Preliminary Analyses

In this section we describe the items used in our analyses, and the preparation of analytic variables.

a) Proportions of respondents who self-reported that they would be willing to do a series of tasks for a given incentive level, and average willingness score.

We asked respondents 20 questions about the willingness to collaborate with Netquest beyond answering survey questions. The different activities proposed were classified a priori in different groups:

- Passive measurement on devices they already use, e.g., “Use the accelerometer on your smartphone to measure your physical activity and report it (passively) to Netquest”.
- Wearing special devices to passively monitor activity, e.g., “Wear a small device on your wrist that measures your alcohol consumption and directly sends the information to Netquest”.
- Providing respondents with measurement devices and then having them self-report the results (i.e., they could see the results, and decide to edit their answers),

[2] The final dataset used is available from the first author upon request.

e.g., “Measure your blood cholesterol level using a finger prick we will provide you and self-report the results to Netquest”.

- The provision of physical specimens or bodily fluids, e.g., “Measure your saliva cortisol by chewing special gum for 30 seconds, then putting it in a vial and mailing it to Netquest”.
- Others, e.g., “Let your children answer surveys that we would send to you for them”.

Our goal was to vary the requests on several dimensions including the frequency of measurement (one time versus continuous), the degree of respondent involvement (passive versus active), the sensitivity of the topic (blood alcohol levels versus photos of products), and so on. The full list of items appears in Table 1.

These questions were separated in two sets of 10 questions each, which differed on two additional levels:

- The incentive offered in exchange for collaboration with the request: 30 points in the first set versus 40 points in the second. These points can be exchanged for gifts by the panelists: for instance, with 20 points, a panelist can get an e-book; with 40 points, an online film; with 120 points, a cinema ticket. Across 186 surveys administered to Netquest panelists in Spain in 2016, the number of points received per survey varied from four to 58 with an average of 14 and a median of 12 (Revilla, 2017).
- The number of answer categories: the first set uses partially labeled 5-point scales from “1- Definitely not” to “5- Definitely yes” whereas the second set uses partially labeled 11-point scales with similar labels (on 0 and 10). A “not applicable” (NA) option was also available.

Note that the incentive and response scales are confounded: the first set used a 5-point scale and 30-point incentive, while the second used an 11-point scale and offered 40 points. We address this confounding later. While we did not randomize items across the incentive and response scale conditions, the order of the items within each set of 10 questions was randomized across respondents in order to minimize potential order effects.

For each question, we look at the proportion of missing answers (respondents were not required to answer each item), the proportion of not applicable (NA) answers, the proportion of respondents who would accept the task (i.e., they answered 4 or 5 for the first set of questions and 6 to 10 for the second set of questions) among those who provided an answer different from NA, and the average willingness rating among those providing an answer different from NA (on a 0-10 scale; transforming the score for the first set by subtracting 1 and multiplying by 2.5; following Preston and Colman, 2000, and Dawes, 2008).

b) Self-reported reasons for accepting or not accepting installation of a browser tracker.

For these analyses, we use the answers to two open questions[3]. The first asked about the reasons why respondents said they would accept (or not) the invitation to install an application on their PC which registers the URLs of the websites they visit and report this (passively) to Netquest.

The second, asked only to respondents who said they would not be willing to install the tracking application or who selected the middle answer category, was: "What would Netquest most need to change such that you would accept the invitation to install an application on your PC which registers the URLs of the websites you visit and report this (passively) to Netquest?"

The answers to these questions were coded by a native Spanish speaker. When a respondent provided several reasons, we consider them all.

c) Factor analysis to identify common elements in willingness to participate in different tasks.

Next, we study what affects willingness to participate in different tasks in a more general way. We expected the tasks proposed in these questions to pertain to different categories of activities (see subsection a). In order to empirically examine the grouping of these tasks, we conducted a principal component factor analysis (PCA) based on the 817 respondents who provided a substantive answer (excluding NA) to all items[4]. Three factors with an eigenvalue greater than 1 were identified. Given that we expected the three factors to be correlated, we considered a 3-factor solution with oblique rotation.

The PCA suggested the following classification of tasks. For a full description of the items, we refer to Table 1.

- Factor 1 ("*PhysicalMeasures*") includes six items about sharing different physical measures: *PassiveStress, CholesterolSelfReport, AlcoholSelfReport, PassiveAlcohol, CholesterolVial, CortisolVial.*
- Factor 2 ("*BehaviorTracking*") includes six items about allowing the fieldwork company to track behavior: *PassiveGPS, TrackerPC*, *TrackerMobile, FacebookProfile, Emotion, EyeMovement.*
- Factor 3 ("*RespondentControl*") includes four items where the respondent has control on the reporting: *PhotosProduct*, *ScanBarcodes, TestProduct, PhotosMobile.*

[3] To reduce burden, we asked the open questions only about one selected task.

[4] Running the PCA on the larger sample with 17 items (excluding the three items about children where the NA proportions are high) yields essentially the same factor structure.

The final four items were not classified because they loaded on two factors equally (*Accelerometer* and *ChildrenStress*) or had low loadings on all factors (*ChildrenSurvey* and *ChildrenWeight*). Based on this classification, we created a willingness score for each of the three factors identified above, using the procedure described below.

First, for the items in the first set, we recoded the answers from 0 to 4 instead of 1 to 5 and multiplied this by 2.5 to get a score from 0 to 10 (as we have for the items in the second set). Then, for each of the three factors identified, we averaged the rescaled items to get an equally-weighted willingness score from 0-10 for each factor. We did not use factor scores because of the varying levels of missing data across items. We excluded those respondents who did not provide substantive answers to at least half of the items in the factor. This means that 7.1% of respondents did not get a summary score for factor 1, while 7.4% did not get one for factor 2, and 6.2% for factor 3.

d) Regression analyses of the willingness to participate on independent variables related to trust, personality and respondent socio-demographics.

The scores on these three factors form the key dependent variables in our analyses. The correlations obtained between the respective factor scores are as follows: 0.62 between factors 1 and 2, 0.63 between factors 1 and 3 and 0.52 between factors 2 and 3. Given the relatively high correlations, we also consider an overall willingness score computed on all 20 items[5]. An examination of the standardized normal probability plots of the summated scales suggests that they approximate normal distributions, justifying the use of OLS regression.

In addition, since the literature does not systematically identify factors affecting willingness (most of the studies are simply descriptive, reporting the rates of willingness or compliance), we selected independent variables which we expected to be associated with these three factors (Appendix A provides details on all the variables and scales):

- Some basic socio-demographic variables: *Men*, *Age*, *Education* and *Income*.
- One question on the frequency of Internet use on a smartphone (*InternetFrequency*). The more frequently respondents use a smartphone to connect to Internet, the more they are likely to use GPS, social media, etc., that already capture this information. Thus we expect they will be more willing to share data of different kinds too.
- Three variables related to the sharing of content (*ShareFB, ShareTwitter, LikeSharingLife*). The more respondents already share content on Facebook and

[5] We also created a score based on the 17 items that do not involve children, and obtain equivalent results.

Twitter, and the more they like sharing their personal life, the more we expect that they will be willing to provide Netquest with different kinds of data.

- Three questions about the benefits of market research (*BenefitForMe, BenefitConsumers* and *BenefitSociety*). The more respondents value market research, the more we expect they will be willing to participate in different tasks.
- Three questions related to trust (*Suspicious*, *SocialTrust*, *TrustAnonymity*). The more trust people have in general and in the anonymity of the information they share, the more we expect them to be willing to share different kinds of data.
- Two questions about the attitude toward safety (*SecureSurroundings* and *AvoidRisk*). The more respondents are concerned about safety in general, the more we expect they will also be worried about the risks of sharing different data with a panel company.
- Three variables related to the attitude toward answering surveys (*AnswerIncome*, *LikeAnswering*, *PriorParticipation*). The more positive respondents' attitudes toward answering surveys (i.e., provided an answer to the income question, liked answering the current survey, answered many previous surveys in the panel), the more we expect them to be willing to participate in other tasks too.
- One question about the attitude toward new activities (*LikeNew*). The more respondents are looking for new things to do, the more we expect them to accept new tasks.

Several of the questions described above (*LikeSharingLife, Suspicious, SocialTrust, SecureSurroundings, AvoidRisk, LikeAnswering,* and *LikeNew*) were part of a separate experiment on agree-disagree (AD) versus item-specific (IS) wording. In a series of models (not shown) we tested whether the different formats affected the relationship of these variables with the willingness factor scores. We tested both main effect models (with an indicator for AD/IS) and interactions (of format with items). Our general conclusion was that the format in which these questions were asked did not have an effect on the conclusions drawn from the models, except for two variables (*SocialTrust* and *LikeNew*). Given that these two variables also did not show consistent or significant relationships with willingness, we decided to drop them from the model. For parsimony, for the other questions, we combine the alternative versions and use them as predictors in the models below.

We examined the bivariate associations between all these variables and the scores created for total willingness and for the three factors *PhysicalMeasures*, *BehaviorTracking*, and *RespondentControl*, and fitted a variety of models. We decided to drop two more variables: *Income* because of the high proportions of missing data (24.7% missing or "I prefer not to answer") and *InternetFrequency* because it did not have a significant effect in the models and we had variables more directly related to the sharing of content through social media (*ShareFB* and *ShareTwitter*).

In addition, we also used the same set of variables to estimate three structural equation models (SEM). In each case, our dependent variable is a latent variable, measured by the different items identified as forming one of the three willingness factors. In terms of independent variables, *Men*, *Age* and *Education* are measured with a single indicator each, whereas the others are measured with several indicators: *Share* is measured with three indicators (*ShareFB*, *ShareTwitter* and *LikeSharingLife*), as is *Benefit* (*BenefitForMe*, *BenefitConsumers*, *BenefitSociety*), and *Attitude toward surveys* (*AnswerIncome*, *LikeAnswering*, *PriorParticipation*), whereas *Trust* and *Safety* are measured with two indicators each (respectively, *Suspicious* and *TrustAnonymity*; and *SecureSurroundings* and *AvoidRisk*). The model was estimated in LISREL and tested using global fit measures as well as the JRule software (Van der Veld, Saris and Satorra, 2009). The model is corrected step by step until an acceptable fit is obtained. Appendix B provides an example of the path diagram for the initial model, as well as a list of the extra parameters introduced in each model in order to get an acceptable fit, and the final estimates of the parameters in each model. Only a summary of the main effects of the SEM are presented in the results section.

Main Results

Stated Willingness to Complete Different Tasks in Exchange for Specific Incentives

We first examine the responses to the 20 individual willingness items. Table 1 provides for each item, the percentage not answering that item (% missing), the percentage of NA answers among those who gave an answer, the percentage who say they would accept the task, and the average willingness score (among those who gave an answer different from NA), ranked by percent willing.

Table 1 Stated willingness, ordered by proportions of accepting (highest to lowest)

If you would receive 30 (or 40*) points in exchange, would you accept the invitation to...	% missing	% NA	% would accept	Average (0-10 scale)
... receive a product at home to test and report on it in a survey (*TestProduct*)*	5.3	1.7	73.7	7.4
... take photos of products with your smartphone and send them to Netquest (*PhotosProduct*)	6.2	2.7	56.4	6.2
... scan barcodes of products with your smartphone and share them with Netquest (*ScanBarcodes*)	5.9	3.5	53.7	6.0

Table 1 continued

If you would receive 30 (or 40*) points in exchange, would you accept the invitation to...	% missing	% NA	% would accept	Average (0-10 scale)
... measure the amount of alcohol in your breath using a breathalyzer test kit we would provide you and self-report it to Netquest (*AlcoholSelfReport*)	6.3	5.8	51.0	5.5
... take photos with your smartphone and send them to Netquest (*PhotosMobile*)*	5.7	2.8	49.6	5.3
... wear a small device on your wrist that measures your stress and directly sends the information to Netquest (*PassiveStress*)	5.8	3.2	44.8	5.0
... measure your blood cholesterol level using a finger prick we will provide you and self-report the results to Netquest (*CholesterolSelfReport*)	6.1	3.3	40.5	4.5
... wear a small device on your wrist that measures your alcohol consumption and directly sends the information to Netquest (*PassiveAlcohol*)*	5.5	5.2	37.8	4.1
... use the accelerometer on your smartphone to measure your physical activity and report it (passively) to Netquest (*Accelerometer*)	6.8	4.6	37.4	4.6
... let your children answer surveys that we would send to you for them (*ChildrenSurvey*)	6.2	25.8	30.7	3.7
... measure your blood cholesterol level using a finger prick we will provide you, then putting it in a vial and mailing it to Netquest (*CholesterolVial*)*	6.0	2.9	30.2	3.3
... measure your saliva cortisol by chewing special gum for 30 seconds, then putting it in a vial and mailing it to Netquest (*CortisolVial*)*	5.9	3.2	27.7	3.1
... measure your children‘s weight when we ask you and self-report it to Netquest (*ChildrenWeight*)*	5.8	25.3	27.5	3.2
... share GPS information from your smartphone with Netquest (*PassiveGPS*)	6.0	3.8	20.8	2.7
… let us record your face while you watch a video in your PC in order to measure the movement of your eyes (*EyeMovement*)*	6.1	3.3	19.3	2.3
... give Netquest access to all the information of your profile on Facebook (as if they were one of your friends) (*FacebookProfile*)*	5.6	5.7	19.0	2.4
… let us record your face while you watch a video in your PC in order to measure your emotional response (*Emotion*)*	5.1	3.6	18.0	2.2

Table 1 continued

If you would receive 30 (or 40*) points in exchange, would you accept the invitation to...	% missing	% NA	% would accept	Average (0-10 scale)
... install an application on your smartphone which register the URLs of the websites you visit and report this (passively) to Netquest (*TrackerMobile*)	6.4	4.3	17.8	2.4
... install an application on your PC which register the URLs of the websites you visit and report this (passively) to Netquest (*TrackerPC*)	6.2	5.0	16.6	2.3
... let your children wear a small device on their wrist that measures their stress and directly sends the information to Netquest (*ChildrenStress*)*	5.5	24.4	11.8	1.5

Note: Tasks followed by a * correspond to the second set (40 points incentive). N = 1,476 for the % missing; N varies from 1,375 to 1,400 for the % NA and from 1,028 to 1,374 for the % would accept and average scores. The % would accept and average columns are based on those who gave a substantive answer (i.e., excluding both the missing and NA responses).

The levels of item missing values are quite similar (ranging from 5.1% to 6.8%). Concerning the levels of NA, the three items asking about children clearly differ from the others, which is to be expected since some of the panelists do not have children[6].

The proportions of respondents willing to accept the different tasks show large variations. The most accepted task is that of receiving a product at home to test and report on in a survey: 73.7% of respondents who gave an answer said they would be willing to do this. This is followed by taking photos of products with a smartphone (already much lower: 56.4%). At the other extreme, the task with the lowest level of willingness consists of letting one's children wear a small device on their wrist that measures their stress and directly sends the information to Netquest, with only 11.8% expressing willingness.

It is interesting to note that the willingness to use a breathalyzer and self-report the readings (51.0%) and the willingness to measure one's blood cholesterol level and self-report the results (40.5%) are much higher than the willingness to give Netquest access to all the information in one's Facebook profile (19.0%) or to install an application on one's PC to register the URLs of websites visited and report this (passively) to Netquest (16.6%).

It is also interesting that there is a difference of more than 10 percentage points between the stated willingness for doing a cholesterol test depending if the results

[6] However, including or excluding the NA answers to compute the willingness has little effect on the rank order and conclusions overall, even for these three items.

are self-reported by the respondents or if the test is directly sent to Netquest via mail. In the second case, where respondents cannot change the results or decide whether or not to share them, stated willingness is lower. Similarly, a difference of 13.2 percentage points is seen in the case of alcohol tests, depending on whether the results are self-reported or directly sent to the panel company. This suggests that respondents make a distinction both on the types of data being measured and on the degree of control they have over what is captured or reported.

Similar results can be seen when considering the average score instead of the proportion of respondents willing to complete the tasks.

In terms of the effect of the differential incentive and response scale, the ordering of items in Table 1 suggests there is not a strong differential effect of the incentive offered. In fact, the average willingness score for the 10 items in the first set (30 points and 5-point scale) is higher than that for the second set (40 points and 10-point scale), likely reflecting differences in the tasks being asked about more than differences in incentives or response scale. This again suggests we can ignore these confounding factors.

Overall, the mean for the total score of willingness is 4.0 (on a 0-10 scale). Considering the three factors, *RespondentControl* has the highest mean (6.2), followed by *PhysicalMeasures* (4.2) and finally *BehaviorTracking* (2.4).

Self-reported Reasons for Being Willing or Not

Next, we focus on one of the tasks proposed in the first set of questions: the willingness to install a passive browser tracking application on one's PC, for which only 16.6% of the respondents who gave an answer expressed willingness to do so. Table 2 reports the main reasons mentioned in a follow-up open question about why they would accept or not the invitation to install a tracking application.

The main reason mentioned for accepting the task was that respondents did not mind or did not feel that this was confidential information (37.4%), followed by interest in getting the incentive (25.1%), altruism (14.0%) and trust (9.9%). On the other side, the main reason for not accepting this task is linked to privacy concerns (72.6%), with a further 7% raising issues of trust.

In order to improve the acceptance of this task, respondents who said that they would not be willing to install the tracking application or chose the middle category were asked what could be done to help them change their decision. While 68.0% of the respondents said that there is nothing that could be done to make them change their mind, 11.7% mentioned improvements in security and 9.7% increased incentives. Even if for a large majority of respondents, it seems unlikely they will be convinced to install a tracking application on their PC, security and incentives are aspects which could improve the overall acceptance of such tasks.

Table 2 Main reasons* why panelists would accept or not accept the invitation to install a tracking application on their PC

Main reasons for accepting	% (based on N= 171 respondents)
I don't mind/not confidential	37.4
Incentive	25.1
Altruism	14.0
Trust	9.9
Main reasons for not accepting	**% (based on N= 829 respondents)**
Privacy	72.6
No trust	7.0
No reason	5.4
I do not own the PC I use	5.8

Note: * We present all reasons that are mentioned by at least 5% of the respondents. When a respondent provided several reasons, we take them all into account.

Predictors of Willingness to Accept Additional Tasks

Table 3 contains a set of four linear regression models, first predicting the total score for all 20 willingness items, and then predicting each of the three scores on the factors identified earlier. In each case a positive coefficient means greater willingness. The overall proportion of variance explained by the set of predictors ranges from 0.22 (for *RespondentControl*) to 0.35 (for the full 20-item scale).

We also run a series of partial F-tests to see if the collective contribution of each group of variables (we have five groups in the final models, besides the socio-demographic variables: attitude toward sharing, benefit of market research, trust, attitude toward safety and toward surveys) was significant, when explaining each of our four dependent variables of interest. All the tests indicated a statistically significant contribution except one: the test for attitude toward safety in the case of the factor *RespondentControl* (F(2, 1039)=2.73; p=.0657). This suggests that each set of variables has an association with willingness.

Several variables are statistically significant across all four models, with coefficients in a consistent direction. As expected, the more frequently respondents report posting content on Facebook, the more willing they are to accept a variety of additional research tasks. *ShareTwitter* is not statistically significant in any of the models, but this may be because fewer respondents use Twitter relative to Facebook (57.2% versus 90.1%), or that Twitter is a more public social networking service. Similarly, *LikeSharingLife* is positively associated with willingness. Those who

Table 3 Regression analyses

		TotalScore		*Physical Measures*		*Behavior Tracking*		*Respondent Control*	
Explanatory variables		Coef.	p-value	Coef.	p-value	Coef.	p-value	Coef.	p-value
Demo-graphics	Men	.342	**.019**	.475	**.018**	.463	**.003**	-.073	.692
	Age	-.004	.521	.006	.474	-.001	.909	-.019	**.019**
	Education	-.165	**.035**	-.302	**.005**	-.123	.141	-.010	.920
Share	ShareFB	.131	**.000**	.153	**.002**	.119	**.002**	.141	**.002**
	ShareTwitter	-.018	.618	-.056	.259	.064	.100	-.028	.543
	LikeSharingLife	.400	**.000**	.400	**.000**	.464	**.000**	.285	**.002**
Benefit	BenefitForMe	.136	.139	.103	.408	.206	**.034**	-.004	.973
	BenefitConsumers	.248	**.013**	.425	**.002**	.110	.298	.332	**.008**
	BenefitSociety	.101	.296	-.062	.642	.205	**.048**	.141	.251
Trust	Suspicious	.054	.451	.012	.904	.102	.185	.018	.843
	TrustAnonymity	.607	**.000**	.645	**.000**	.584	**.000**	.402	**.006**
Safety	SecureSurroundings	.058	.402	.113	.232	-.078	.291	.202	**.022**
	AvoidRisk	-.139	**.008**	-.236	**.001**	-.088	.113	-.087	.190
Attitude Toward Surveys	AnswerIncome	.537	**.002**	.509	**.030**	.582	**.001**	.506	**.018**
	LikeAnswering	1.232	**.000**	1.242	**.000**	1.108	**.000**	1.334	**.000**
	PriorParticipation	.277	**.000**	.250	**.008**	.306	**.000**	.208	**.016**
	Constant	-4.333	**.000**	-3.561	**.002**	-5.891	**.000**	-2.475	**.020**
	No. observations	1,044		1,052		1,049		1,056	
	R-squared	.345		.237		.330		.216	
	Adj. R-squared	.335		.225		.320		.204	

Note: coefficients in bold when statistically significant (p-value<.050)

have greater trust in the anonymity of their data are more willing to accept additional tasks. Those who answered the income question (indicating a degree of trust or willingness to disclose) also show higher levels of willingness on all four measures. Finally, two indicators of survey engagement are positively associated with willingness: those who liked answering the survey and those who have responded to more prior Netquest surveys have higher levels of willingness.

Several other variables are statistically significant in some but not all of the models. The effect of gender is statistically significant (men more willing) for three of the four models. The coefficient for education is negative (those with higher education less willing) for all four models but only reaches statistical significance for the *TotalScore* and *PhysicalMeasures* models. Those who perceive greater benefit of research for consumers have significantly higher willingness for three of the four measures (*TotalScore*, *PhysicalMeasures*, and *BehaviorTracking*), but the direction of the effect is consistent across all four models. Those who are inclined to avoid risk have significantly lower levels of willingness on *TotalScore* and *PhysicalMeasures,* but not on the other two factors (although, again, the effect is in a consistent direction).

Finally a few variables reach statistical significance in only one of the models. Age has a significant negative effect (older people less willing) only for the *RespondentControl* factor. Both those who see a personal benefit and those who see a societal benefit of market research are more willing to agree to *BehaviorTracking*. Finally, those who rate *SecureSurroundings* as more important are *more* willing to agree to tasks that permit respondent control.

Overall, we see largely consistent effects of predictors across the different types of activities, although there is enough variation among the models to suggest that different types of people react differently to the different types of additional tasks being asked about.

Considering the SEM analyses, quite similar results are obtained, even if there are few differences. The latent variables *Share*, *Benefit* and *AttitudeTowardSurveys* have statistically significant positive effects on the three willingness factors. In addition, men have higher willingness on *PhysicalMeasures* and *BehaviorTracking*. Finally, *Safety* has a statistically significant negative effect on *BehaviorTracking* (the more one cares about safety, the less willing). Details of the SEM are presented in Appendix B.

Discussion

In this paper we investigated the willingness to perform additional tasks among panelists of an opt-in online panel in Spain. We found that the willingness to perform additional tasks is not a unitary phenomenon. Respondents distinguish between different types of tasks, and are more willing to do some but not others. In general, willingness is higher for tasks where respondents have control over the reporting of the results (e.g., taking pictures, measuring one's blood cholesterol level and reporting the results) than for passive tracking behaviors (e.g., installing a tracking app on one's PC or smartphone), even if this means that respondents have to do more work than with passive measurements where they only need to give their permission once. This is probably due to high privacy concerns, which is also what the answers to the open questions suggest: most respondents mentioned reasons related to the issue of trust/security/privacy both for accepting or not accepting the installation of a tracking app.

Our factor analysis revealed three distinct but related types of tasks: *PhysicalMeasures*, *BehaviorTracking*, and *RespondentControl*. Our models also suggest that there are variables that reliably predict willingness, as measured by these factors. This implies that restricting a sample to only those willing to accept a specific task is likely to result in both demographic and attitudinal biases.

The study has several limitations. The results are based on an opt-in panel (already generally cooperative, self-selected, already have a relationship with the panel), and further restricted to those who have Internet access through both a PC and a smartphone. We are studying stated willingness, not actual willingness. The results are restricted to a single panel (Netquest) in a single country (Spain). Thus we should be cautious about generalizing the results to different panels and countries. There are also some limitations in the analyses performed: some questions were asked in different formats (AD versus IS) for random subsets of respondents; some answered the survey on a PC, others on a smartphone (again, randomly assigned). Also we could not really take the difference in incentives into account (i.e., we did not randomly assign respondents to different incentive conditions); however, our primary focus was not on the incentives but the tasks.

In addition, we identified a variety of different tasks, but did not systematically try to vary the features or elements of these tasks, such as the degree of intrusiveness, the potential burden, the degree of respondent control, etc. More work is needed to explore the various dimensions that affect willingness to perform some tasks but not others. Our research has started looking at the "what" (i.e., what people are willing to do and not do) but not as much at the "why" (why people are willing or not, although our open question started to address this issue). More research is needed to explore the reasons behind differential willingness of panelists to accept different tasks, and to understand how stated willingness translates to actual compliance.

Researchers are increasingly exploiting the measurement capabilities of modern technologies. Understanding how consumers react to these requests, and understanding the differences between those who are willing and those who are not, are important steps in evaluating the utility of these additional tasks or measures. Most of the prior studies have focused only on a single task (e.g., GPS capture, or installing a browser tracker). Our research finds that treating all tasks as the same, and making inference from one type of request to all other requests, is risky. The stated willingness to use new technologies to provide additional data to researchers varies according to the nature of the task. A first step to overcoming the barriers to accepting new technologies is understanding within- and between-respondent differences in willingness.

References

Angrisani, M., Kapteyn, A., & Samek, S. (2017). Real time measurement of household electronic financial transactions in a population representative panel. Paper presented at the ESRA conference, Lisbon, July.

Antoun, C., Couper, M.P., & Conrad, F.G. (2017). Effects of mobile versus PC web on survey response quality: A crossover experiment in a probability web panel. Public Opinion Quarterly, 81(5), 280-306.

Armoogum, J., Roux, S., & Pham, T.H.T. (2013). Total nonresponse of a GPS-based travel surveys. Paper presented at the conference on New Techniques and Technologies for Statistics, Brussels, March.

Avendano, M., Scherpenzeel, A., & Mackenbach, J.P. (2011). Can biomarkers be collected in an Internet survey? A pilot study in the LISS panel. In M. Das, P. Ester, & L. Kaczmirek (Eds.), Social Research and the Internet (pp. 371-412). New York: Taylor and Francis.

Biler, S., Šenk, P., & Winklerová, L. (2013). Willingness of individuals to participate in a travel behavior survey using GPS devices. Paper presented at the conference on New Techniques and Technologies for Statistics, Brussels, March.

Boase, J. (2016). Augmenting survey and experimental designs with digital trace data. Communication Methods and Measures, 10(2-3), 165-166.

Boase, J, & Ling, R., (2013). Measuring mobile phone use: Self-report versus log data. Journal of Computer-Mediated Communication, 18(4), 508-519.

Boyle, J., Kilpatrick, D., Acinerno, R., Ruggiero, K., Resnick, H., Galea, S., Koenan, K., & Galernter, J. (2010). Biological specimen collection in an RDD telephone survey: 2004 Florida hurricanes gene and environment study. In L.A. Aday & M. Cynamon (Eds.), Ninth Conference on Health Survey Research Methods (pp. 176-184). Hyattsville, MD: National Center for Health Statistics.

Burton, J. (2016). Results for Web/face-to-Face linkage consent questions in the Innovation Panel. Paper presented at the Mixing Modes and Measurement Methods in Longitudinal Studies Workshop. London: CLOSER.

Couper, M.P., & Singer, E. (2013). Informed consent for web paradata use. Survey Research Methods, 7(1), 57-67.

Couper, M.P., Singer, E., Conrad, F.G., & Groves, R.M. (2008). Risk of disclosure, perceptions of risk, and concerns about privacy and confidentiality as factors in survey participation. Journal of Official Statistics, 24(2), 255-275.

Couper, M.P., Singer, E., Conrad, F.G., & Groves, R.M. (2010). An experimental study of disclosure risk, disclosure harm, incentives, and survey participation. Journal of Official Statistics, 26(2), 287-300.

Crawford, S.D., McClain, C., Young, R.H., & Nelson, T.F. (2013). Understanding mobility: Consent and capture of geolocation data in web surveys. Paper presented at the annual meeting of the American Association for Public Opinion Research, Boston, May.

Dawes, J. (2008). Do data characteristics change according to the number of scale points used? An experiment using 5-point, 7-point and 10-point scales. International Journal of Market Research, 50 (1), 61-77.

De Reuver, M., & Bouwman, H. (2015). Dealing with self-report bias in mobile Internet acceptance and usage studies. Information & Management, 52(3), 287-294.

Etter, J.-F., & Bullen, C. (2011). Saliva cotitine levels in users of electronic cigarettes. European Respiratory Journal, 38(5), 1219-1220.

Gatny, H.H., Couper, M.P., & Axinn, W.G. (2013). New strategies for biosample collection in population-based social research. Social Science Research, 42, 1402-1409.

Gautier, A., Rahib, D., Brouard, C., Saboni, L., Blineau, V., El Malti, F., David, C., Chevaliez, S., Barin, F., Larsen, C., Lot, F., & Lydién N. (2016). Proposition d'un volet biologique à l'issue d'une enquête téléphonique: retour d'expérience du BaroTest. 9ème colloque francophone sur les sondages, Gatineau, octobre 2016. Available at: http://sondages2016.sfds.asso.fr/programme/programme-avec-presentations/

Gilbert, E., Conolly, A., Tietz, S., Calderwood, L., & Rose, N. (2017). Measuring young people's physical activity using accelerometers in the UK Millennium Cohort Study. London: Centre for Longitudinal Studies, CLS working paper 2017/15.

Greenfield, T.K., Bond, J., & Kerr, W.C. (2014). Biomonitoring for improving alcohol consumption surveys: The new gold standard? Alcohol Research: Current Reviews, 36(1), 39-45.

Hassani, M., Kivimaki, M., Elbaz, A., Shipley, M., Singh-Manoux, A., et al. (2014). Nonconsent to a wrist-worn accelerometer in older adults: the role of socio-demographic, behavioural and health factors. PLoS ONE, 9 (10), e110816.

Howie, E.K., & Straker, L.M. (2016). Rates of attrition, non-compliance and missingness in randomized controlled trials of child physical activity interventions using accelerometers: a brief methodological review. Journal of Science and Medicine in Sport, 19, 830–836.

Jäckle, A., Burton, J., Couper, M.P., & Lessof, C. (2017). Participation in a mobile app survey to collect expenditure data as part of a large-scale probability household panel: response rates and response biases. Institute for Social and Economic Research, University of Essex: Understanding Society Working Paper Series No. 2017-09.

Joh, K. (2017). 2017-2018 regional household travel survey. Presentation to the National Capital Region Transportation Planning Board, Travel Forecasting Subcommittee, May 19th.

Johnson, A., Kelly, F., & Stevens, S. (2012). Modular survey design for mobile devices. Paper presented at the CASRO Online Research Conference, Las Vegas, February.

Keusch, F., Antoun, C., Couper, M.P., Kreuter, F., & Struminskaya, B. (2017). Willingness to participate in passive mobile data collection. Paper presented at the annual meeting of the American Association for Public Opinion Research, New Orleans, LA, May.

Kissau, K., & Fischer, D. (2016). Pitfalls and opportunities of research using passive metering software. Paper presented at the General Online Research conference, Dresden, March.

Lauderdale, D.S., Schumm, P.L., Kurina, L.M., McClintock, M., Thisted, R.A., Chen, J.H., & Waite, L. (2014). Assessment of sleep in the National Social Life, Health, and Aging Project. Journals of Gerontology, Series B: Psychological Sciences and Social Sciences, 69(8), S125–S133.

Link, M.W., Murphy, J., Schober, M.F., Buskirk, T.D., Childs, J.H., & Tesfaye, C. (2014). Mobile technologies for conducting, augmenting and potentially replacing surveys: Report of the AAPOR task force on emerging technologies in public opinion research. Public Opinion Quarterly, 78(4), 779-787.

McFall, S.L., Conolly, A., & Burton, J. (2014). Collecting biomarkers using trained interviewers. Lessons learned from a pilot study. Survey Research Methods, 8(1), 57-66.

McGeeney, K., & Weisel, R. (2015). App vs. web for surveys of smartphone users experimenting with mobile apps for signal-contingent experience sampling method surveys. Washington, DC: Pew Research Center report, http://www.pewresearch.org/2015/04/01/app-vs-web-for-surveys-of-smartphone-users/

Menai, M., van Hees, V.T., Elbaz, A., Kivimaki, M., Singh-Manoux, A., & Sabiaa, S. (2017). Accelerometer assessed moderate-to-vigorous physical activity and successful ageing: results from the Whitehall II study. Science Reports, 7, 45772; doi: 10.1038/srep45772.

Miller, G. (2012). The smartphone psychology manifesto. Perspectives on Psychological Science, 7(3), 221-237.

Palmer, J.R.B., Espenshade, T.J., Bartumeus, F., & Chung, C.Y. (2013). New approaches to human mobility: Using mobile phones for demographic research. Demography, 50(3), 1105-1128.

Pinter, R. (2015), Willingness of online access panel members to participate in smartphone application-based research. In D. Toninelli, R. Pinter, & P. de Pedraza (Eds.), Mobile Research Methods: Opportunities and Challenges of Mobile Research Methodologies (pp. 141-156). London: Ubiquity Press.

Preston, C.C., & Coleman, A. (2000). Optimal number of response categories in rating scales: reliability, validity, discriminating power, and respondent preferences. Acta Psychologica, 104, 1-15.

Raento, M., Oulasvirta, A., & Eagle, N. (2009). Smartphones: An emerging tool for social scientists. Sociological Methods & Research, 37, 426-54.

Revilla, M. (2017). Analyzing survey characteristics, participation, and evaluation across 186 surveys in an online opt-in panel in Spain. Methods, Data, and Analysis, 11(2), 135-162.

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

Revilla, M., Ochoa, C., & Loewe, G. (2017). Using passive data from a meter to complement survey data in order to study online behavior. Social Science Computer Review, 35(4), 521-536.

Roth, A., & Mindell, J.S. (2013). Who provides accelerometry data? Correlates of adherence to wearing an accelerometry motion sensor: the 2008 Health Study for England. Journal of Physical Activity and Health, 10, 70-78.

Sakshaug, J.W., Couper, M.P., & Ofstedal, M.B. (2010). Characteristics of physical measurement consent in a population-based survey of older adults. Medical Care, 48(1), 64-71.

Scherpenzeel, A. (2017). Mixing online panel data collection with innovative methods. In S. Eifler & F. Faulbaum (Eds.), Methodische Probleme von Mixed-Mode-Ansätzen in der Umfrageforschung (pp. 27-49). Wiesbaden: Springer.

Toepoel, V., & Lugtig, P. (2014). What happens if you offer a mobile option to your web panel? Evidence from a probability-based panel of Internet users. Social Science Computer Review, 32(4), 544-560.

Van der Veld, W.M., Saris, W.E., & Satorra, A. (2009). Jrule 2.0: User Manual. Radboud University Nijmegen, The Netherlands.

Van Duivenvoorde, S., & Dillon, A. (2015). The best of both worlds? Combining passive data with survey data, its opportunities, challenges and upside. Paper presented at the CASRO Digital Research Conference, February 11-12, Nashville, TN.

Wenz, A., Jäckle, A., & Couper, M.P. (2017). Willingness to use mobile technologies for data collection in a probability household panel. Institute for Social and Economic Research, University of Essex: Understanding Society Working Paper Series No. 2017-10.

Wrzus, C., & Mehl, M.R. (2015). Lab and/or field? Measuring personality processes and their social consequences. European Journal of Personality, 29, 250–271.

Appendix A

List of all independent variables considered: exact formulation and scales

- *Men* (1="Male", 0="Female")
- *Age* (in years)
- *Education* (in six categories, from no education to university degree)
- *Income* (in six categories, from lowest to highest)
- *Internet Frequency*: "On average, how frequently do you connect to the Internet using a smartphone" ("1=Once a month or less" to "6=Daily")
- Sharing of content:
  - *Share FB*: "In general, how frequently do you share content on your personal Facebook account?" (8 response options ranging from "I don't have a personal account" to "I share content every day")
  - *Share Twitter*: "In general, how frequently do you share content on your personal Twitter account?" (same 8 response options)
  - *Like sharing life*: either asked in an agree-disagree format ("I like sharing my personal life", 1="Completely disagree" to 5="Completely agree") or an item-specific format ("How much do you like sharing your private life?" 1="Don't like at all" to 5="Like extremely").
- Benefits of market research for:
  - the respondent him/herself (*Benefit for me*)
  - consumers (*Benefit consumers*)
  - the society/citizens (*Benefit society*)

  (1="Does not benefit at all" to 5="Benefits a great deal").
- Trust:
  - *Suspicious*: "I get suspicious easily" (1="Completely disagree" to 5="Completely agree") or "How easily do you get suspicious?" (1="Not at all easily" to 5=" Extremely easily")
  - *Social Trust*: "I don't trust people in general" (1="Completely agree" to 7="Completely disagree") or "How much do you trust people in general?" (1="Do not trust at all" to 7="Trust completely")
  - *Trust anonymity*: "To what extent do you trust that this survey guarantees anonymity?" (1="Do not trust at all" to 4="Trust completely")

- Attitude toward safety:
    - *Secure surroundings*: “It is important for me to live in secure surroundings” (1=“Completely disagree” to 7=“Completely agree”) or “How important is it for you to live in secure surroundings?” (1=“Not important at all” to 7=“ Extremely important”)
    - *Avoid risk*: “I always avoid anything that can endanger my safety” (1=“Completely disagree to 7=“Completely agree) or “How often do you avoid anything that can endanger your safety?” (1=“Never” to 7=“Always”)

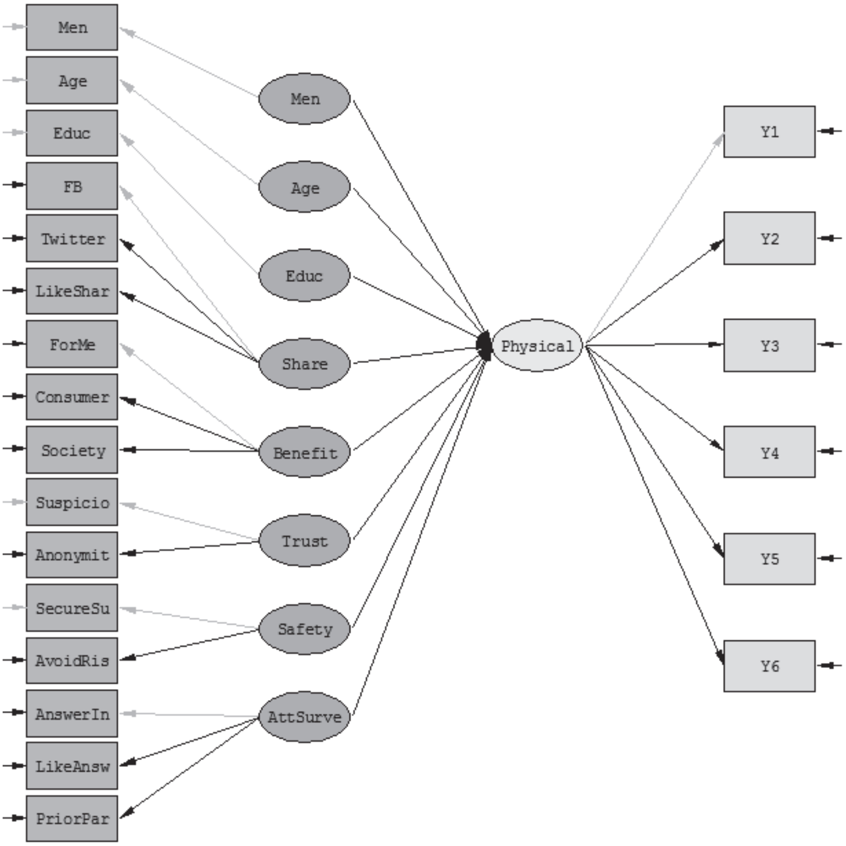
- Attitude toward answering surveys:
    - *Answer Income:* dummy variable coded 1 if the respondent provided a substantive answer to the income question, and 0 otherwise (no answer at all, or “I prefer not to answer” option).
    - *Like Answering*: “How much did you like or not to fill in this questionnaire? (1=“Did not like it at all” to 4=“Liked it very much”)
    - *Prior participation*: number of Netquest surveys completed before this one, recoded from lowest to highest into quartiles.

- Attitude toward new activities:
    - *Like New:* “I am never looking for new things to do” (1=“Completely agree” to 5=“Completely disagree”) or “How often are you looking for new things to do?” (1=“Never” to 7=“Always”)

Appendix B

More information about the SEM analyses

a) Initial model for factor 1 (LISREL path diagram); models are similar for factors 2&3



b) Extra parameters introduced in each model in order to get an acceptable fit

Model Factor 1: correlated error terms for Y5 and Y6; Y2 and Y5; Age and Prior Participation; and cross-loading Att. Survey and Anonymity.

Model Factor 2: correlated error terms for Y5 and Y6; Y2 and Y3; Age and Prior Participation; and cross-loading Att. Survey and Anonymity.

Model Factor 3: correlated error terms for Age and Prior Participation; and cross-loading Att. Survey and Anonymity.

c) Estimates of the parameters in each model (completely standardized solution).

		Physical Measures	Tracking Behavior	Respondent Control
Measurement model	F by Y1	.79 ^{NA}	.80 ^{NA}	.88 ^{NA}
	F by Y2	.77*	.76*	.81*
	F by Y3	.79*	.72*	.64*
	F by Y4	.88*	.71*	.85*
	F by Y5	.75*	.66*	Not present
	F by Y6	.73*	.66*	Not present
	Share by FB	.76NA	.71 ^{NA}	.75 ^{NA}
	Share by Twitter	.49*	.52*	.50*
	Share by Like Sharing	.39*	.45*	.39*
	Benefit by For Me	.84 ^{NA}	.85 ^{NA}	.84 ^{NA}
	Benefit by Consumer	.86*	.86*	.86*
	Benefit by Society	.82*	.82*	.82*
	Trust by Suspicious	1.00 ^{NA}	1.00 ^{NA}	1.00 ^{NA}
	Trust by Anonymity	-.14*	-.14*	-.14*
	Safety by Secure Surroundings	1.00 ^{NA}	1.00 ^{NA}	1.00 ^{NA}
	Safety by Avoid Risk	.42*	.43*	.43*
	Att. Survey by Anonymity	.49*	.51*	.50*
	Att. Survey by Answer Income	.26 ^{NA}	.26 ^{NA}	.26 ^{NA}
	Att. Survey by Like Answering	.64*	.60*	.67*
	Att. Survey by Prior Participation	.22*	.24*	.20*
Structural model	Men on F	.07*	.11*	-.03
	Age on F	.02	.01	-.05
	Education on F	-.09*	-.05	-.02
	Share on F	.15*	.30*	.16*
	Benefit on F	.17*	.19*	.17*
	Trust on F	.00	.02	-.01
	Safety on F	.02	-.06*	.04
	Att. Survey on F	.51*	.60*	.49*
Fit	Chi-Square	$\chi^2(202)=742.94$	$\chi^2(202)=728.30$	$\chi^2(165)=672.81$
	RMSEA	.053	.052	.056

Note: F refers to the factor of interest (*PhysicalMeasures* or *TrackingBehavior* or *RespondentControl*). Y1 to Y6 refer to the items used to measure this F factor (thus there are different in each model). * Indicates a coefficient statistically significantly different from 0 (t-ratio >1.96). ^{NA} indicates that no values are available for the t-ratio because the corresponding loading was fixed to 1 (unstandardized) for identification purposes.

methods, data, analyses | Vol. 13(2), 2019, pp. 253-266 DOI: 10.12758/mda.2018.02

The Advantage and Disadvantage of Implicitly Stratified Sampling

Peter Lynn

Institute for Social and Economic Research, University of Essex

Abstract

Explicitly stratified sampling (ESS) and implicitly stratified sampling (ISS) are well-established alternative methods for controlling the distribution of a survey sample in terms of variables that define the strata. If these variables are correlated with survey estimates, the estimates will benefit from improved precision. With ESS, unbiased estimation of the standard errors of survey estimates is possible, provided that sampling strata membership is identified on the survey dataset. With ISS this is not possible and usual practice is to invoke an approximation that tends to result in systematic over-estimation of standard errors. This can be perceived as a disadvantage of ISS. However, this article demonstrates, both theoretically and through a simulation study, that true standard errors can be smaller with ISS and argues that this advantage may be more important than the ability to obtain unbiased estimates of the standard errors. The simulation findings also suggest that the extent of over-estimation with the usual approximate variance estimator may be modest.

Keywords: standard error, stratified sampling, survey sampling

Most surveys use stratified sampling designs. This is done in order to benefit from the precision gains that such designs can bring. For a modest effort in designing the sample, the precision gains can often be equivalent to those that would accrue from carrying out tens or even hundreds of extra interviews. Stratified sampling is therefore highly cost-effective. However, there are many different ways that it can be done. The researcher must choose which variables to use, and how to combine them to define the strata. She must also decide whether all strata should be sampled at the same rate (proportionate stratified sampling) or whether some should be over-sampled, perhaps in order to increase the representation in the sample of certain subgroups (disproportionate stratified sampling). Though the researcher is typically constrained to define strata in terms of information that is either available on the sampling frame or can be linked to the frame, this still usually leaves a lot of options regarding exactly how the information should be used. The better the decisions, the more cost-effective the survey design will be.

This article focuses on one specific decision that the researcher must make: whether to use explicitly stratified sampling (ESS) or implicitly stratified sampling (ISS). For simplicity, the arguments are illustrated in the context of proportionate stratified sampling, but the arguments apply equally when sampling is disproportionate, as a similar decision must be made within each top-level sampling domain. The arguments also apply when a decision is being made about how to stratify at a secondary level, i.e. within primary explicit strata.

ESS involves sorting the population elements into explicit groups (strata) and then selecting a sample independently from each stratum. ISS involves ranking the elements following some ordering principle and then applying systematic sampling, i.e. selecting every n^{th} element. For example, if the sampling frame were a list of people containing a single auxiliary variable, date of birth, proportionate ESS would involve creating strata corresponding to a number of discrete age groups and then selecting, using simple random sampling (SRS), a number of people from each group such that the proportion of the sample in each group equals the proportion of the population in the group. ISS, on the other hand, would involve sorting the people from youngest to oldest (or oldest to youngest; this is equivalent) and then selecting every n^{th} person on the list (after generating a random start point).

The advantage of ESS is that unbiased estimation of the standard errors of survey estimates is possible, provided that the sampling stratum membership is identified on the survey dataset and provided that at least two sample elements are selected from each stratum. With ISS this is not possible and usual practice is to

Acknowledgments

The author is grateful to Seppo Laaksonen for provoking him to write this article.

Direct correspondence to

Peter Lynn, ISER, University of Essex, Wivenhoe Park, Colchester,
Essex CO4 3SQ, UK
E-mail: plynn@essex.ac.uk

invoke an approximation that tends to result in systematic over-estimation of standard errors. This can be perceived as a disadvantage of ISS. However, this begs the question of whether it is better to know the precision of one's estimates or to have more precise estimates without knowing exactly how much more precise they are.

On the other hand, there are several disadvantages of ESS relative to ISS. One of these relates to the focus of this article: a greater precision gain due to stratified sampling can be achieved with ISS than with ESS (Madow and Madow, 1944; Cochran, 1946). Another disadvantage of ESS is that it is not possible to obtain an equal-probability sample unless each stratum size is an exact multiple of the sampling interval. Consequently, unequal design weights must be applied, with an associated further loss in precision. Furthermore, it is not possible to stratify deeply on a combination of many variables, due to restricting limitations on the number of strata and an associated risk of greater variation in the design weights the larger the number of explicit strata relative to the sample size. Deeper stratification is possible with ISS.

ESS is often used in order for different sampling fractions to be applied to different sub-domains of the population (disproportionate stratified sampling), by creating the strata to reflect the sub-domains. However, this should not be perceived as an advantage of ESS as the same can be achieved with ISS by assigning a size measure to each element proportional to the desired sampling fraction and making selections with probability proportional to this size measure. Variance estimation for variable probability systematic sampling is considered by Stehman and Overton (1994).

The potential of ISS to provide a greater precision gain than ESS is recognised in the statistical literature (e.g. Madow & Madow, 1944; Kish, 1965) but is not given attention in the sample design sections of generalist survey research handbooks or textbooks. For example, Groves et al. (2009) explain ESS and how, with proportionate allocation to strata, it can improve precision compared to simple random sampling (pp. 113-120). They then introduce systematic selection as "a simpler way to implement stratified sampling" (p. 122), but make no mention of the implications for precision, other than a rather general statement that "Systematic sampling from an ordered list is sometimes termed "implicitly stratified sampling" because it gives approximately the equivalent of a stratified proportionately allocated sample" (p. 124). Even texts that are devoted specifically to sampling, when written for non-statisticians, do not mention explicitly how ESS and ISS compare in terms of precision. For example, Henry (1990) states that "Systematic sampling has statistical properties that are similar to simple random sampling" (p. 98), and subsequently, "Another advantage is that systematic sampling can be used for de facto stratification to insure proportional representation of the population for some characteristic" (p. 98), but with no further mention of precision. In a subsequent section on ESS, however, Henry states that "stratification reduces standard errors" (p. 101) and demonstrates how this works with formulas and a worked example. Kalton (1983) too explains the variance properties of ESS at some length (pp. 20-24), while the shorter section devoted to ISS focusses instead on the practicality of implementation: “systematic sampling provides a mean of substantially reducing the effort required for sample selection” (p. 16).

Even the most recent specialist texts on survey sampling provide very little detail on the statistical properties of ISS. Bethlehem (2009) merely points out that, “the sample variance […] need not necessarily be a good indicator of the variance of the estimator” (p. 79) and then suggests that the only way to obtain an unbiased estimator for the variance is to select multiple samples and combine the observed sample means. Approximations are not mentioned. Valliant et al (2013) state that, “Systematic sampling is often used in practice because it is fairly easy to implement and it can be used to control the distribution of a sample across a combination of auxiliary variables” (p. 63) and “Regardless of the reasons for its use, statisticians usually collapse the selection intervals into one or more analytic strata and pretend the method of selection was something else, like *stsrswor*, *stsrswr*, or *ppswr*, in order to estimate a variance.” (p. 64). It is therefore unsurprising if survey researchers may have the impression that ESS is the (only) way to improve precision compared to SRS.

Furthermore, empirical demonstrations of the relative performance of ESS and ISS are surprisingly hard to find. This article provides an exposition of the distinction between ESS and ISS and attempts, via a simulation study using real survey data, to quantify the extent of the improvement in precision with ISS and the extent of the uncertainty about the improvement in precision if the usual approximation is used to estimate standard errors. In the next section, the relevant aspects of sampling theory are presented and are used to derive an expression for the difference in sampling variance between ESS and ISS. The subsequent sections describe how a simulation study will be used to quantify the true difference in sampling variance between the two designs and the extent to which sampling variance will tend to be over-estimated if the usual approximation is used in the case of ISS. The results from the study are then presented and the implications are discussed in the final section.

Sample Designs and Variance Estimators

For simplicity of exposition, it will be assumed that survey estimates are means or proportions. Under ESS, the sampling variance of the sample mean can be expressed (Kish, 1965, p. 81; Cochran, 1977, p. 69) as:

$$Var\left(\overline{y}\right)=\sum_{i=1}^{I}\frac{N_i S_i^2\left(N_i-n_i\right)}{\left(N^2 n_i\right)} \tag{1}$$

where $S_i^2 = Var_i\left(y_{ik}\right)$ is the variance of y within stratum i (y_{ik} is the value of y for individual k in stratum i);

n_i is the number of sample elements in stratum i;

N_i is the number of population elements in stratum *i*;

and $N = \sum_{i=1}^{I} N_i$ is the total number of elements in the population.

In this article we will assume the context of proportionate sampling, in which case $\frac{n_i}{N_i} = \frac{n}{N}, i = 1,\ldots,I.$ With this assumption, expression (1) simplifies to:

$$Var\left(\bar{y}\right) = \frac{\sum_{i=1}^{I} S_i^2 \left(N_i - n_i\right)}{nN} \tag{2}$$

From this expression it can be seen that differences between strata in terms of *y* do not contribute to the sampling variance. The sampling variance depends only on the variance of *y* within the strata. This demonstrates how stratified sampling improves the precision of estimates; by eliminating any influence on the sample of one part of the variance of *y*, namely the part that is between-strata. Once a survey has been carried out, assuming equal probabilities of selection, $Var\left(\hat{y}\right)$ can be estimated in a straight-forward manner from the survey data, by substituting the observed within-stratum sample variances $\left(s_i^2\right)$ for the corresponding population variances $\left(S_i^2\right)$, thus:

$$\widehat{Var}\left(\bar{y}\right) = \frac{\sum_{i=1}^{I} s_i^2 \left(N_i - n_i\right)}{nN} \tag{3}$$

For ISS designs there is of course no concept of explicit strata, so the $\{i\}$ in expression (2) are not defined. The design-based variance of a sample mean is equivalent to that under cluster sampling with a sample size of one cluster (Madow & Madow, 1944). Unbiased sample-based estimators of this variance do not exist. While a number of estimators have been proposed, all of them are biased and all will over-estimate the variance whenever the stratification effect is anything more than negligible (Wolter, 1984; Wolter, 1985, pp. 258-262). A commonly-used variance estimation method is to treat the ordered list of selected elements as if each consecutive pair had been selected from the same stratum, a method referred to by Kish (1965, p. 119) as the “paired selections model”, and by Wolter (1985, pp. 250-251) as the “estimator based on nonoverlapping differences”. Thus, a systematic sample of *n* elements from an implicitly-stratified list is treated as if it consisted of simple random samples of size 2 from each of *n*/2 explicit strata. Analogous methods, in which elements selected from more than one stratum are treated as if they had been selected from the same stratum, are also sometimes used in the context of ESS, particularly when there exists one or more strata in which only one element is selected or observed (Cochran, 1977; Seth, 1966; Rust & Kalton, 1987). In order to compare the sampling variance of ISS and ESS, we can consider the situation in which the ISS pseudo-strata are subsets of the ESS strata. This is a realistic reflection of the example mentioned in the previous section of stratifying either explicitly or implicitly using date of birth. We will denote the ISS substrata by $j = 1, \dots , J_i$. Then, the approximation usually invoked to estimate the sampling variance associated with ISS is:

$$\widehat{Var}\left(\bar{y}\right) = \frac{\sum_{i=1}^{I}\sum_{i=1}^{J_i} s_j^2\left(N_j - n_j\right)}{nN} \tag{4}$$

whereas the true ISS sampling variance is:

$$Var\left(\bar{y}\right) = \frac{\sum_{h=1}^{N/n}\left(\bar{y}_h - \underline{\underline{y}}\right)^2}{\left(N/n - 1\right)} \tag{5}$$

where

there are *N/n* possible samples that could be selected, corresponding to the *N/n* possible random start points;

$\bar{y}_h$ is the sample mean of *y* for sample *h*;

$\underline{\underline{y}} = \frac{n}{N}\sum_{h=1}^{N/n} \bar{y}_h$ is the mean of the *N/n* sample means.

This true variance can be thought of as the sampling variance of a mean under cluster sampling, with a sample size of one cluster, where the population is divided into *N/n* clusters, $\bar{y}_h$ are the cluster means, and $\underline{\underline{y}}$ is the population mean.

Expression (4) is also used as an estimator for 1-per-stratum designs. In this case, the estimator is known to be upwardly-biased (Fuller, 2009, p. 202; Breidt et al., 2016). ISS is similar to 1-per-stratum sampling, so the bias in using expression (4) as an estimator for (5) might be assumed to be similar, but the designs are not exactly equivalent. In particular, with ISS stratum boundaries are arbitrary and are constrained only conditionally on the random start, and ordering within strata is not random. It should be clear from expression (4) that both ISS and 1-per-stratum ESS should provide greater precision than the most precise form of ESS that enables unbiased estimation of standard errors, namely 2-per-stratum ESS $\left(\sum_{i=1}^{I} J_i = {}^{n}\!/_{2}\right)$. If the ordering of elements within each stratum in a 2-per-stratum design is completely random, then further sub-dividing each stratum *j* into two substrata ($k_j = 1,2$) to create a 1-per-stratum design will have no effect on the sampling variance as $s_{k_j}^2 = s_j^2 \forall k_j$. But any meaningful ordering of elements within at least some of the strata will result in $s_{k_j}^2 < s_j^2$ for at least some *j*, *k*, and hence reduced sampling variance. Wolter (1985) presents a series of simulations in which the estimator based on nonoverlapping differences is shown to sometimes be upwardly-biased and sometimes downwardly-biased as an estimator of the ISS variance, depending on the nature of the population ordering.

Simulation Methodology

Data from wave 1 of *Understanding Society, the UK Household Longitudinal Study*, are treated as population data. These data are used to calculate the sampling variance of means and proportions under simple random sampling, ESS and ISS, in ways that will be described in this section. *Understanding Society* is a large nationally-representative multi-topic general population survey. A stratified, multi-stage sample of addresses was selected (Lynn, 2009) and all persons aged 16 or over resident at a sample address were eligible for an individual interview at wave 1. Members of ethnic minority groups and residents of Northern Ireland were sampled at higher rates than the remainder of the population. Data collection took place face-to-face in respondent's homes using computer-assisted personal interviewing (CAPI) between January 2009 and March 2011. At wave 1, 50,295 individual interviews were completed with sample members. For the illustrative purposes of this article, these individuals are treated as a population from which survey samples are to be selected.

A set of eleven target parameters were selected for study. Of these, five are means of continuous variables and six are proportions based on binary variables. For each, we are interested in comparing the sampling variance of the sample statistic under alternative sampling designs and the estimate of the ISS sampling variance using the successive pairing approach. For ease of exposition and calculation, for each parameter we first amend the population such that N is a multiple of 100. This allows the subsequent creation of equal-sized explicit strata (each containing $N_i = 100$ elements) and the application of implicitly stratified systematic sampling designs in which the sampling interval takes the integer value of 50, the convenience of which will be explained below. From the 50,295 elements, we first drop any with item missing values. This is done separately for each of the eleven target variables, so the dropped elements will differ between the eleven simulated populations. Then, a further set of m elements are dropped (m between 0 and 99) in order to round the population size down to a multiple of 100. The m elements with the smallest analysis weights (largest inclusion probabilities) are chosen. Descriptive statistics regarding this process are presented in Table 1.

For each estimate, the variance and estimated variance for samples of size $N/50$ will be compared under different designs. These designs are simpler than those that tend to be used for real social surveys. Specifically they are all equal-probability single-stage designs, without clustering, and with stratification based on a single auxiliary variable, whereas real designs often involve variable probabilities, multi-stage selection, clustering and multiple stratification variables. The simplifications are introduced in order to provide a simple illustration in which differences between the designs are strictly limited to the aspects of design that are the focus of this article. The following sub-sections describe the sampling variance metrics that were calculated for each of the eleven parameters to be estimated. All but one of the metrics rely on knowledge of the population size, N, and the popula-

Table 1 Simulated Populations for 11 Parameter Estimates

	Understanding Society sample size	Item missing	Also dropped (smallest weights)	Simulated population size, N	Sample size, n
Continuous variables					
Total monthly income	50,295	78	17	50,200	1,004
Monthly benefit income	50,295	3,236	59	47,000	940
Number of children	50,295	50	45	50,200	1,004
Hours of sleep	50,295	12,420	75	37,800	756
Body mass index	50,295	6,432	63	43,800	876
Binary variables					
Limiting long-term illness (%)	50,295	0	95	50,200	1,004
Arthritis (%)	50,295	3,234	61	47,000	940
In paid employment (%)	50,295	90	5	50,200	1,004
Has degree (%)	50,295	86	9	50,200	1,004
Lives with spouse/partner (%)	50,295	0	95	50,200	1,004
Religion makes a great difference (%)	50,295	3,234	61	47,000	940

Note: Hours of sleep was asked in a supplemental self-completion questionnaire that was returned by only 85.9% of interview respondents, whereas all other items were administered in the face-to-face interview. The items on body mass index, arthritis and religion were not included in the proxy version of the face-to-face interview, which was administered for 6.4% of respondents.

tion variance of *y*, S^2, each of which were derived in the usual way from the population simulated as described above.

Simple Random Sampling

The variance of $\overline{y}$ under simple random sampling is computed as a benchmark and will be used later in the calculation of design effects for the various sample designs under consideration, to help with interpretation of the findings. It is calculated in the usual way:

$$Var_{SRS}\left(\overline{y}\right)=\frac{S^2\left(N-n\right)}{nN} \qquad (6)$$

Explicit Stratified Sampling with 11 Strata

The first stratified design considered is one with eleven explicit strata, defined by the person's age. The first stratum consists of persons aged 16 to 19; the following nine strata consist of five-year age bands from 20-24 to 60-64; the final stratum consists of person 65 years old or older. Proportionate stratified sampling with a sampling fraction of 1 in 50 is used. The sampling variance of a mean is therefore calculated as in expression (2) above, with $n_i = {}^{N_i}\!\big/_{50}$ and $I = 11$.

Explicit Stratified Sampling with *N*/100 Strata

The second stratified design considered is one with N/100 equal-sized explicit strata, again defined by the person's age. It can be seen from Table 1 that this corresponds to between 378 and 502 strata. The strata are created by first sorting the population in increasing order of age and then treating the first 100 in sorted order as the first stratum, and so on. A simple random sample of n = 2 is selected from each stratum. The sampling variance of a mean is therefore calculated as in expression (2) above, with $n_i = 2$ and $I = N/100$.

Implicit Stratified Sampling with *n* = *N*/50

The third design considered involves sorting the population in increasing order of age and then selecting a systematic random sample of *N*/50 cases using a random start between 1 and *N*/50. There are therefore *N*/50 possible samples that could be selected and the sampling variance of a mean is calculated as the variance of the *N*/50 corresponding sample means, as in expression (5), with $n = 50$.

In addition to calculating the true sampling variance for this design, the expected value of the estimated sampling variance was calculated using the consecutive pairs method outlined in section 2 above. This was done by calculating the estimate produced by expression (5) for each of the *N*/50 possible samples and then taking the mean of these *N*/50 values.

Results

For each of the eleven variables, Table 2 presents the true standard error of the sample mean under each of the four sample designs under consideration, as well as the expected value of the estimate of the standard error for the ISS design under the consecutive pairs method. The true value of the population mean is also presented for reference (first column). It is worth noting firstly that the relative standard errors vary greatly between the eleven estimates. Under SRS, they range from 0.01 to

0.08, with the exception of body mass index, which has a relative standard error of 0.65 (driven by a number of influential outliers). This provides a range of circumstances in which to compare the effects of alternative stratified sample designs.

As expected, standard errors are in all cases smaller under stratified sampling than under simple random sampling. In fact the rank order of the four designs in terms of standard error is the same for all eleven estimates: ESS with eleven strata provides an improvement in precision over SRS, ESS with around 500 strata (*N*/100) provides a further improvement, and ISS improves precision further still. The relative extent of the standard error reduction varies between the estimates, however. For example, for estimating mean number of children or the proportion of people in paid employment most of the gains to be had from stratifying by age accrue with the use of just eleven explicit strata: extensions to 500 strata or ISS provide only very modest marginal gains. For body mass index and for the proportion suffering from arthritis, on the other hand, the gains in moving from eleven to 500 explicit strata are similar or greater in magnitude to those in moving from no strata (SRS) to eleven. These differences evidently reflect the differing nature of the associations of the variables with age and are illustrated in Figure 1, which presents the design effect for each of the three stratified designs (ratio of sampling variance under ESS or ISS to that under SRS). The proportion suffering from arthritis stands out as the estimate that gains most in terms of precision from each of the successive enhancements to stratification. The precision gain in moving from the ESS11 to the ESS(N/100) design demonstrates that tendency to suffer from arthritis is quite strongly associated with age, even within the eleven strata of the ESS11 design. However, the further gain in moving to the ISS design shows that even within (at least some of) the 470 strata in the ESS(N/100) design there remains an association of arthritis with age. This may seem surprising considering that each of the 470 strata covers an age range of only around 2.5 months, on average, but is explained by the strata towards the upper end of the age range – where arthritis is most prevalent – covering larger age ranges, reflecting the smaller population sizes. The design effect of around 0.65 for this estimate with ISS – the smallest of all the design effects in this study – represents a very considerable precision gain. Without stratification, this improvement in precision would require an increase in the sample size with SRS from 940 to 1,443 – an increase that would have considerable cost.

The other variable that stands out in Figure 1 is the only attitudinal variable in the study, the proportion of people agreeing with the statement that religion makes a big difference in life. This variable stands out because the precision gains from stratification are much more modest than for all other variables. Beliefs about the importance of religion are only very weakly associated with age.

Turning now to the final column of Table 2, it can be seen that the consecutive pairs method of variance estimation for ISS results in a modest over-estimation of standard errors, i.e. an under-estimation of the precision gain from stratification. The expected value of the estimated standard error is typically similar to, or just slightly smaller than, the true standard error with the ESS(N/100) design. This is of course the design that is assumed by expression (4) with $n_j = 2$, but the estimated standard errors differ from the true standard errors under this design due to the data having been generated by a different mechanism.

Table 2 Standard errors of means and proportions under four sample designs, and mean estimated standard errors for implicit stratified sampling

		s.e.				Est.(s.e.)
	Mean	SRS	ESS(11)	ESS(N/100)	ISS	ISS
Continuous variables						
Total monthly income	1479.0	49.82	47.28	46.22	45.33	46.24
Monthly benefit income	466.0	37.28	36.23	36.08	35.72	36.15
Number of children	1.600	0.0467	0.0406	0.0403	0.0403	0.0403
Hours of sleep	6.97	0.0587	0.0577	0.0576	0.0574	0.0576
Body mass index	26.06	17.03	16.42	15.35	15.19	15.28
Binary variables						
Limiting long-term illness (%)	34.93	1.489	1.400	1.397	1.392	1.396
Arthritis (%)	14.29	1.130	1.042	0.976	0.912	0.968
In paid employment (%)	52.29	1.560	1.362	1.339	1.333	1.339
Has degree (%)	21.37	1.281	1.242	1.236	1.216	1.226
Lives with spouse/partner (%)	61.51	1.520	1.384	1.356	1.338	1.344
Religion makes a great difference (%)	22.13	1.340	1.337	1.334	1.331	1.333

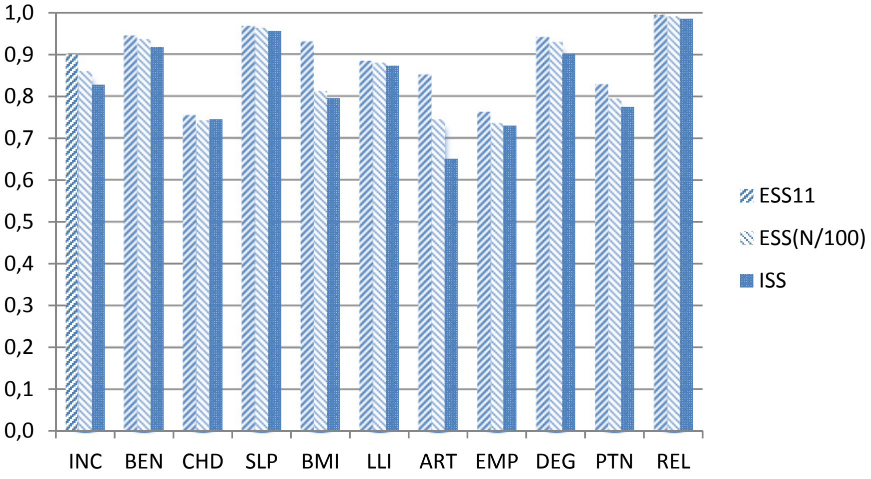


Figure 1 Design effects for three sample designs

Discussion

The simulation study has shown, using real survey data, that ISS provides useful precision gains relative to ESS. This is true even when comparing to the most detailed form of ESS possible, namely that which involves creating strata such that just two selections are made from each stratum (i.e. the minimum number that permits variance estimation.) This result should lead researchers to question why, whenever useful auxiliary data are available for sample stratification, one would ever choose not to use implicit stratification, given that estimates will be less precise as a result. In practice, ESS typically involves a rather smaller number of strata, such that the average number of sample elements selected from each stratum is very considerably greater than two, perhaps more akin to the ESS11 design presented here, in which around 90 elements are selected per stratum. In this study, the ISS design produced substantially smaller standard errors than the ESS11 design. Gains are apparent, though more modest, even relative to the ESS(N/100) design. There consequently seems to be a strong case for ISS designs rather than ESS designs of this kind.

Furthermore, the approximation commonly used to estimate standard errors with ISS results in only a modest over-estimation. This would make statistical tests slightly conservative, which is probably more desirable than the false precision that would be provided by the opposite. In any case, the extent of the over-estimation (systematic error) is most likely small compared to the extent of sampling variance in the standard error estimate (random error). This conclusion is consistent with that of Wolter (1985, p. 283) who compared eight different possible variance estimators for systematic sampling and concluded that the consecutive pairs estimator "performed, on average, as well as any of the estimators" and "in very small samples … might be the preferred estimator".

The choice between ESS and ISS would therefore seem to come down to a choice between improved precision of the survey estimate or unbiased estimation of the precision of the survey estimate. To take the estimation of the proportion of people suffering from arthritis as a concrete example, would researchers prefer to have a standard error of 0.976 associated with their estimate (expected value) of 14.29 (the smallest standard error that would be possible with ESS) and to have an estimate of the standard error with an expected value of 0.976, or to have a standard error of 0.912 (with ISS) and an estimate of the standard error with an expected value of 0.968? For descriptive estimation, it is hard to imagine why the less precise estimate might be preferred. The choice could be less clear, however, when the objective is statistical inference. Analysts could justifiably prefer unbiased hypothesis tests, including those that are implicit in the fitting of statistical models. This distinction between different kinds of analysis objectives is particularly problematic for surveys that are used for both types of analysis, as only one sample design can be used. The ideal solution might be to develop ways of adjusting in inferential analysis for the bias in the variance estimator.

It should be noted that results could be different if a combination of multiple stratification variables were used rather than a single variable, as in the simulations presented here. With a single stratification variable, it is likely that any relationship of the implicitly stratified ordering with the target parameters will be monotonic, or at most quadratic in nature, whereas when combining variables large discontinuities in the distribution can occur at the boundaries of categories of a variable. However, there is no suggestion in Wolter (1985, p.268) that the bias in the consecutive pairs estimator is strongly dependent on whether one, two or three stratification variables are used.

A limitation of the empirical results presented here is that they are restricted to full-sample means and proportions. Some additional simulations (results not shown) for subclass means and proportions based on the same variables suggest that ISS less frequently provides a noticeable improvement in precision over the ESS(N/100) design. This could be because relatively few of the strata in the ESS(N/100) design provide more than one element in the subclass, in which case there is little scope for further precision gains. However, to explore this limitation further, analysis should be extended to a range of subclasses, with different distributions over strata, and to other types of ratio estimates. Such investigation is beyond the scope of this article.

A final point to note is that the situation considered here is that of single-stage sample selection. In practice, stratification is also sometimes used at one or more stages of a multi-stage design. For example, many address-based surveys use stratification at the first stage but not at the final stage (e.g. Lynn, 2009; Lynn & Lievesley, 1991). The precision gains due to stratification are generally likely to be more modest in such designs than in single-stage designs, and consequently the differences between ISS and ESS may also be more modest. A different situation is where stratification is used at the final stage of a multi-stage design. An example might be the selection of pupils within schools after first selecting a sample of schools. In this situation, precision gains can be considerable and it seems likely that the effects described in this article should apply. Indeed, the likely small sample size within each primary sampling unit is likely to result in ISS having even greater advantages, for the design weight reasons discussed in section 1 above.

References

Breidt, F.J., Opsomer, J.D., & Sanchez-Borrego, I. (2016). Nonparametric variance estimation under fine stratification: an alternative to collapsed strata. *Journal of the American Statistical Association*, 111:514, 822-833.

Cochran, W. (1946) Relative accuracy of systematic and stratified random samples for a certain class of populations. *The Annals of Mathematical Statistics*, 17(2), 164-177.

Cochran, W. (1977). *Sampling Techniques*, 3rd Edition. New York: John Wiley.

Fuller, Wayne A. (2009). *Sampling Statistics.* Hoboken, NJ: Wiley.

Groves, R.M., Fowler, F.J., Couper, M.P., Lepkowski, J.M., Singer, E., & Tourangeau, R. (2009). *Survey Methodology* (2nd edition). Hoboken, NJ: Wiley.

Henry, G.T. (1990). *Practical Sampling*. Newbury Park, California: Sage.

Kalton, G. (1983). *Introduction to Survey Sampling.* Sage Quantitative Applications in the Social Sciences Series, paper 35, Beverly Hills, California: Sage.

Kish, L. (1965). *Survey Sampling.* New York: John Wiley.

Lynn, P. (2009). Sample design for Understanding Society. *Understanding Society Working Paper* 2009-01, Colchester: University of Essex.

Lynn, P., & Lievesley, L. (1991). *Drawing General Population Samples in Great Britain*. London: SCPR

Madow, W. G., & Madow, L. G. (1944). On the theory of systematic sampling, I. *Annals of Mathematical Statistics.* 15, 1–24.

Rust, K., & Kalton, G. (1987). Strategies for collapsing strata for variance estimation. *Journal of Official Statistics*. 3, 69-81.

Seth, G. R. (1966). On collapsing of strata. *Journal of the Indian Society of Agricultural Statistics.* 18, 1-3.

Stehman, S.V. & Overton, W.S. (1994). Comparison of variance estimators of the Horvitz-Thompson estimator for randomised variable probability systematic sampling. *Journal of the American Statistical Association*, 89(425), 30-43.

Wolter, K.M. (1984). An investigation of some estimators of variance for systematic sampling. *Journal of the American Statistical Association*, 79(388), 781-790.

Wolter, K. M. (1985). *Introduction to Variance Estimation*. Berlin: Springer-Verlag.

methods, data, analyses | Vol. 13(2), 2019, pp. 267-290 DOI: 10.12758/mda.2017.14

Behavioral Intentions, Actual Behavior and the Role of Personality Traits. Evidence from a Factorial Survey Among Female Labor Market Re-entrants

Katrin Drasch

Friedrich-Alexander University Erlangen-Nürnberg

Abstract

Factorial surveys (FS) are used frequently to draw conclusions about behavior. However, in FS only behavioral intentions are measured and answering fictive situations are likely to be connected with individual personality traits. Therefore, it is unclear to what extent behavioral intentions as measured by FS and actual behavior are related. It is also unclear whether and how personality traits influence intentions and actual behavior. This paper addresses this subject matter by analyzing these research questions. The theory of planned behavior serves as the theoretical basis (Ajzen, 1991).

The research questions are addressed with data from a factorial survey collected among 395 prospective female labor market re-entrants. They were asked about their willingness to accept lower wages if compensated by "positive" nonmonetary job characteristics. A follow-up study after one year also included information on actual behavior, i.e., whether the woman has found a job. The analysis reveals that women who are willing to accept "negative" job characteristics are more likely to re-enter employment, suggesting a high correlation between results from the factorial survey and actual behavior and thus external validity. Furthermore, personality traits only have a minor influence on behavioral intentions and behavior. This confounds previous non-experimental research results. However, some individual effects are different in the intentions and behavioral model, which also indicates differences between experimental and real-world settings.

Keywords: factorial survey, vignette study, personality traits, intentions and actual behavior, mother's labor market re-entry

Factorial surveys (FS) are a powerful tool for collecting information on norms and attitudes (cf. Auspurg & Hinz, 2015). In recent decades, research using FS has rapidly increased (cf. Wallander, 2009). The method makes use of different fictive situations that must be judged in an interview sequentially by the respondents. In addition, FS can also be used to draw conclusions about behavior or more precisely about behavioral intentions (e.g., Abraham et al., 2013; Nisic & Auspurg, 2009).

When measuring behavioral intentions instead of behavior, first the question arises whether behavioral intentions as measured by FS are related to actual behavior. Second, it is unclear how personality traits influence intentions and actual behavior. Third, it is unclear whether the assumed link between intentions and behavior works differently for individuals with different personality traits. Thus, we examine the role of personality traits for the interplay of intentions and behavior. We use the FS framework that allows examining the role of personality traits for the same respondents and studying a similar situation in a fictional as well as real-world setting. This is important because personality traits might affect actual behavior in a different way than they might affect behavioral intentions. In addition, individuals with different personality traits might respond to fictive situations in another kind of way because they are stimulated differently by them. This of course would confute the general applicability of factorial surveys. In sum, this contributes to further knowledge about the external validity of FS which is regarded as a research gap (Auspurg & Hinz, 2015).

So far, research on the comparison of intentions and actual behavior in the FS survey framework has mostly focused on mobility decisions and decision intentions. Nisic and Auspurg (2009) conclude that the intention to move as measured by a factorial survey and realized moves observed in a representative population survey are driven largely by the same factors, although the magnitude of planned and actual moves is different. Also, Hainmueller, Hangartner, and Yamamoto (2015) examine citizenship decisions in a survey experiment and in a behaviorial setting and find that the survey experiment leads to a reliable estimation of the effects as

Acknowledgments

I would like to thank Gerhard Krug for valuable suggestions and comments especially on the modeling approach. Furthermore, I would like to thank participants of the Forschungskolloquium Analytische Soziologie at LMU Munich and participants of the ESRA 2017 conference, Lisbon. I would also like to thank the reviewers of MDA and the editors. All remaining errors are of course mine. I gratefully acknowledge support and data access from the Institute for Employment Research (IAB).

Direct correspondence to

Katrin Drasch, Friedrich-Alexander University Erlangen-Nürnberg, Faculty of Humanities, Social Sciences and Theology, Institute for Sociology, Chair for Methods of Empirical Social Research
E-mail: Katrin.Drasch@fau.de

compared to the real-world. However, little is known about the cognitive processes underlying the response to a factorial survey (Auspurg & Hinz, 2015).

With respect to previous research about the relation between intentions and behavior in general not focusing on the FS framework, psychological research has proven in many different contexts that this relation exists but that the magnitude depends on the specific conditions under study (for an overview we refer to Ajzen, 1991). Psychological research sometimes finds a low correlation between general personality traits and behavior in a specific situation (e.g., Mischel, 1968). In contrast, Back, Schmukle, & Egloff (2009) report that direct and indirect measures of personality predict various types of behavior. However, coming from a survey methodological perspective, we are not interested in studying the intra-individual differences in intentions and behavior but whether (prospective) behavioral intentions as measured with FS and (retrospectively measured) behavior as measured in general social surveys are related to each other and to what extent personality influences this relationship.

Answers of respondents of FS on behavioral intentions and behavior itself can be suspected to be prone to be biased through different personality traits of individuals. Several studies from the field of economics (for an overview we refer to Almlund et al. 2011) have used the concept of personality to study their impact with respect to different labor market behaviors, for example, smoking (Anger, Kvasnicka, & Siedler, 2011) or income (Heineck & Anger, 2010).

With respect to our example - the labor force participation decision of mothers - economic literature has examined the influence of personality traits on actual behavior (Wichert & Pohlmeier, 2010; Berger, 2010). However, this research shows rather mixed results. While Wichert & Pohlmeier (2010, p. 16) conclude that “all personality traits except agreeableness significantly influence the participation decision”, Berger (2010, p. 1) states that “the dimension agreeableness of the Big Five personality traits is found to be associated with later return to employment”. Notably both articles studied the labor force participation of mothers in Germany in a similar timeframe. Both articles are based on the GSOEP data, and the instrument used to measure personality was the Big Five assessment as developed for the SOEP 2005 (BFI-S) (Gerlitz & Schupp, 2005; Dehne & Schupp, 2007). However, what is different is the sample. While Wichert & Pohlmeier (2010) use a cross-sectional dataset, Berger (2010) uses the SOEP as a longitudinal dataset.). Within the context of FS research, the relation between personality traits, behavioral intentions and actual behavior has not yet been studied to our knowledge.

Therefore, this article uses the return decision of mothers who have been out of the labor market for several years to study both behavioral intentions in a FS survey framework and the actual behavior of mothers in a real-world setting. The research questions will be addressed with data from a FS collected among 395 women who are prospective labor market re-entrants. They were asked about their willingness to accept lower wages if compensated by job characteristics that are regarded as more favorable by society (e.g., not overqualified labor). The FS contains information on behavioral intentions that covers the likelihood of accepting a given job offer with certain characteristics. It also contains a short version (15-item version as used in the German Socio-Economic Panel) of the assessment of Five-Factor Model (Big Five) (Dehne & Schupp, 2007). A follow-up study after one year also includes information on actual behavior, i.e., whether a woman has found a job and, if so, the characteristics of this job.

Theoretical Framework and Hypotheses

The theory of planned behavior (TPB) (Ajzen, 1991) as an extension of the theory of reasoned action (Ajzen, 1988) is suitable to derive hypotheses on the influence that personality traits have on both actual and planned behavior. The TBP is a general and parsimonious model that predicts a broard range of behaviors (Connor & Abraham, 2001). According to this theory, attitudes, subjective norms and perceived behavioral control are related to behavior when certain assumptions are met. In our context, this makes it possible to relate re-entry intentions with realized job re-entries of mothers after family-related employment interruptions. Within the TPB also personality as a possible influence factor can be integrated (Connor and Abraham, 2001).

One prerequisite is that the measurement of intentions corresponds to the behavior that is aimed to be predicted (Ajzen, 1991). This is known as the compatibility principle. This principle claims that intentions and actual behavior are closely related when they address the same decision and are measured on the same level. This similarity refers to action, aim, context, and timing (Kalter, 1997). In our example, we measure intentions as re-entry willingness when a specific job offer with certain characteristics is presented. We also examine realized re-entries of mothers. Thus, we relate a decision with restricted information on certain job characteristics to a decision covering most likely more than the described job offer. Similarity with respect to action and aim is thus given. The context, however, is different: while intentions are measured through an experimental setting, realized entries refer to the actual behavior of an individual in a real-world setting. With respect to timing, we conclude that the situation is similar because women who are in the process of re-entry will be examined although they might not yet have been in the situation of being confronted with a job offer when the intention was measured.

Intentions include motivational factors that have an influence on behavior. As such, they are seen as an indicator of the extent to which individuals are willing to exert the actual behavior. Not surprisingly, stronger intentions should lead to a higher likelihood of actually exhibiting the behavior (Ajzen, 1991).

Furthermore, the decision under study must be under the volitional control of the individual. Volitional control refers to whether the person can decide at will to perform or not perform the behavior. What is problematic is that some behaviors meet this requirement better than others. However, when a person has both the opportunity and resources, he or she should also be able to exhibit the behavior. In our example, we look at women who are prepared for a successful labor market re-entry and have the opportunity to accept a given job offer due to the positive general conditions in the German labor market. In sum, we expect that behavioral intentions and actual behavior are closely related *(hypothesis 1)*.

Due to the compatibility principle, rather general personality traits are expected to have no direct influence on the behavior itself *(hypothesis 2)*. Personality traits are assumed to have only an indirect impact by influencing factors that are more closely connected to the behavior (Ajzen, 1991). This is in line with previous argumentations from psychology that traits as broad behavior dispositions are not suitable to be linked with behavior in a very specific situation.

In addition, family (partnership status, age of youngest child) as well as individual characteristics (age, educational attainment, duration of interruption, and place of residence) can be assumed to influence an individual's decision to re-enter the labor market. However, these are not central for our argumentation and we refer to Drasch (2013) for a theoretical elaboration on the effects of those characteristics.

Data and Measurement

Data Collection

Data are taken from a supplement of an evaluation project ('Perspektive Wiedereinstieg' – PWE) developed by the Federal Ministry for Family Affairs, Senior Citizens, Women and Youth (BMFSFJ) and conducted on behalf of the Institute of Employment Research (IAB). This project aimed to re-include women in the labor market who had been inactive for at least three years but want to return to paid employment. In addition, a comparison group consisting of women who have been classified as prospective job returners ("Berufsrückkehrerinnen") not taking part in the evaluation project was generated through matching techniques (NN-matching on the regional level and propensity score matching on the individual level) (Diener et al., 2013). We use both groups and control whether the women belong to one or the other group in the statistical analyses.

A professional social research company conducted CATI interviews with two cohorts of project participants and two comparison groups of registered prospective returners. After the first interview, all women were asked whether they were willing to participate in an add-on online survey containing the FS. If they stated their consent, their e-mail address was then noted down. Thus, the sample under study should be considered as a convenient rather than a representative sample. Knowledge about actual behavior or more precisely whether they had actually re-entered the labor market was generated through wave two panel data.[1]

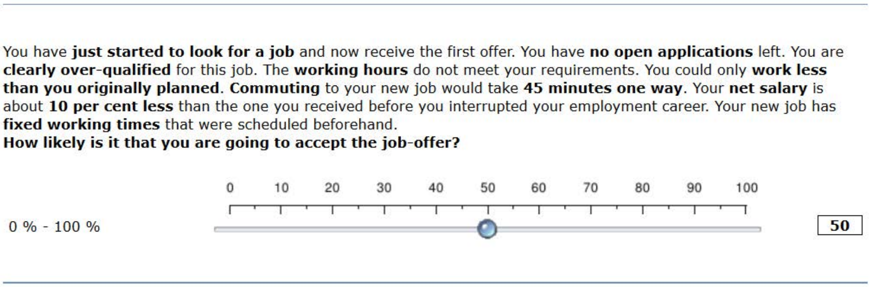
In total, 395 prospective labor market re-entrants can be analyzed with the data. The prospevtive labor market re-entrants were all female because the prerequisite to take part in the program was to have interrupted employment due to family obligations. Because only very few men participated in the program, they were excluded from the quantitative part of the evaluation study. The participating women were asked about their willingness to accept lower wages if they were compensated by more favorable nonmonetary job characteristics. The factorial survey contains information on behavioral intentions, i.e., on the likelihood of accepting a given job offer with certain characteristics, as well as a short version (15 item-version as used in the German Socio-Economic Panel) of the Five-Factor Model (Big Five) that collects information on five central personality traits: neuroticism, extraversion, openness to experience, agreeableness and conscientiousness. Furthermore, we include family as well as individual characteristics as control variables. For more information on the sample characteristics of the FS we refer to Drasch (2013).

Vignette Setup

FS, often alternatively called vignette studies, are suitable to model decisions in complex scenarios (Rossi & Anderson, 1982; Jasso, 2006). Respondents receive several hypothetical scenarios (vignettes) that include an independent variation of a limited number of dimensions. The independence of the dimensions is reached through external variation, which makes a causal interpretation of the dimensions possible. A convenient sample is then sufficient to make predictions about the relevance of the dimensions. Thus, a factorial survey can be regarded as a controlled experiment.

The vignettes consisted of several dimensions that are assumed to have an influence on the re-entry decision, i.e., search phase, search situation, training, work volume, commuting time, wage and working hours. Those dimensions are consistent with previous recommendations on the design of vignettes (cf. Auspurg & Hinz, 2015; Auspurg et al., 2015) of two (search phase and situation) or three (training, volume of work, commuting time, wage and working hours) variations of

[1] Because of data protection regulations of the project, the data and the files cannot be made available to the public.



You have **just started to look for a job** and now receive the first offer. You have **no open applications** left. You are **clearly over-qualified** for this job. The **working hours** do not meet your requirements. You could only **work less than you originally planned**. **Commuting** to your new job would take **45 minutes one way**. Your **net salary** is about **10 per cent less** than the one you received before you interrupted your employment career. Your new job has **fixed working times** that were scheduled beforehand.
How likely is it that you are going to accept the job-offer?

Figure 1 Sample vignette, own translation

the dimensions. These so-called levels were generated bearing in mind meaningful values for the group under study based on a review of the literature on job dimensions. Figure 1 shows a sample vignette.

Answers could be given on a scale ranging from 0 to 100 percent with 5 percent intervals. In sum, 21 answer categories were generated allowing the dependent variable to be treated as metric. More specifically, a number matching technique was used in line with magnitude scaling and the starting point for respondents was set at 0 percent. We are confident that these techniques combine the advantages and disadvantages of both techniques (Schaeffer & Bradburn, 1989).

The 2x2x3x3x3x3x3 levels of the dimensions resulted in 972 possible combinations. None of the combinations had to be excluded due to implausibility. To reduce the number of vignettes to 200, a resolution V design (Dülmer, 2007; Kuhfeld, Randall, & Garratt, 1994; Kuhfeld 2010) was chosen and the levels were orthogonalized to allow for estimation of the main level effects and first order interactions. This resulted in a D-efficient design with a D-efficiency of 98.1 with 100 being the maximum value. This is regarded (cf. Auspurg & Hinz, 2015; Dülmer, 2016) as very efficient.[2] A final step consisted of the random allocation of 10 vignettes to one deck of vignettes resulting in 20 decks. Those decks were then also randomly allocated to the respondents.

Big Five Personality Traits (BFI-S)

The measurement of personality is based on the Big Five approach, which assumes that personality is also reflected in answers to statements about one's attitudes. We use the shortest available two-minute-version for Germany (BFI-S) that covers 15 items measuring the concept's five personality traits: Neuroticism (N), Extraversion (E), Openness to experience (O), Agreeableness (A) and Conscientiousness

[2] We thank Katrin Auspurg, LMU Munich, for the technical implementation in SAS.

Table 1 Dimensions of the BFI-S (translated from German)

Trait	Item I see myself as someone who …	Cronbachs α
extraversion	… is communicative, talkative … is outgoing, sociable … reserved (-)	0.69 (0.61)
agreeableness	… has a forgiving nature … is considerate and kind to others … is sometimes somewhat rude to others (-)	0.50 (0.50)
conscientiousness	… does a thorough job … does things effectively and efficiently … tends to be lazy (-)	0.55 (0.67)
neuroticism	… is relaxed, handles stress well (-) … gets nervous easily … worries a lot	0.64 (0.57)
openness	… is original, comes up with new ideas … has an active imagination … values artistic experiences	0.72 (0.73)

(-) negatively coded items are reversed before analysis; results for Cronbachs α from SOEP 2005 pretest in parentheses

(C). The version was developed for the German Socio-Economic Panel and was used for the 2005 wave (Gerlitz & Schupp, 2005; Dehne & Schupp, 2007). Possible answers were given on a 7-point Likert type scale. One major advantage of this approach is that the descriptive results with respect to the measurement of personality traits can be compared to a general population survey. The items were normalized as described in Dehne and Schupp (2007) and have a mean value of 50. The cronbach's alpha values of the traits range between 0.5 and 0.72. So, the internal reliability of the traits is fairly low but comparable to the values in the GSOEP study. Table 1 shows the Big Five items that were presented in random order to the respondents on an extra page in the online survey.

Willingness to Accept Unfavorable Job Characteristics

To provide a real-world validation and compare the results to the acceptance intentions, we selected a data setup which is displayed in Figure 2.

We restrict our sample to all women who were not employed (including marginally and occasionally employed) when the online factorial survey was con-

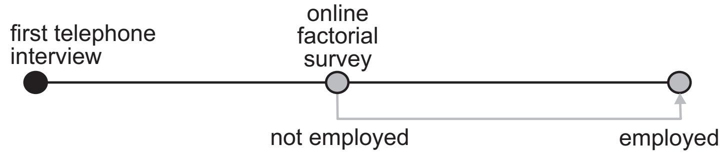


Figure 2 Real-world validation setup

ducted and examine whether they are employed full- or part-time in the follow-up interview about a year later. Thus, we excluded women who were not employed when the first interview was conducted but already reported being employed in the online survey.

As an additional independent variable, we compute a variable that examines the individual deviance (on the vignette level) from the average judgement of the given vignette (without the respondent's own individual judgement to avoid a bias to the average judgement). Thus, the computation of the average judgement on the vignette level is based on around 40 judgments with a range of 25-63 judgements. The following formula illustrates this:

$$Dev_{ij} = X_{ij} - \overline{X}_j \tag{1}$$

This variable displays broadly the individual willingness to accept unfavorable job characteristics as compared to others who are given the same vignette. The standard deviation of this variable amounts to 26 percentage points with a minimum value of -80 and a maximum value of 76, which indicates a large range (see Table A in the appendix). For the empirical analysis, the variable is standardized with a mean value of 0 and a standard deviation of 1.

Empirical Method and Results

Modeling Approach

As the dependent variable for one part of the analyses, we use the vignette judgment (Y). As the set of variables on the vignette level, we use the six job dimensions described above. Furthermore, we include variables on the individual level (Z), including the Big Five personality traits. Age, partnership status, age of youngest child, residence, duration of interruption, and educational attainment were also included as control variables as in Drasch (2013). Table A in the appendix shows the distribution of the independent variables.

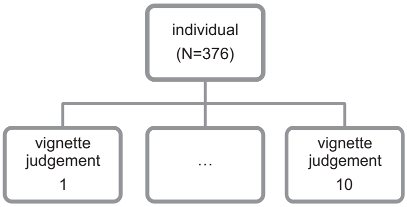


Figure 3a Vignettes nested in individuals

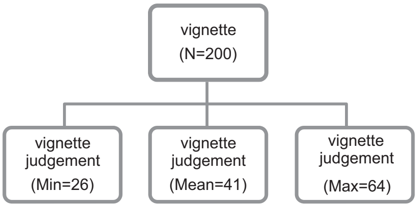


Figure 3b Judgments nested in vignettes

Thus, the data can be considered as multi-level data with two levels. On the superordinate level, the individual is set, and on the subordinate level, the judgment of the ten different vignettes (cf. Figure 3a) is set. However, the data structure becomes more complicated when including the individual deviance and examining realized entries. Then, an alternative approach is to view the vignette as a superordinate level and the different vignette judgements (ranging from 26 to 64 judgements per vignette) made by several individuals as a subordinate level. Figure 3b illustrates this.

Thus, individuals share not only common properties but also vignettes. Ideally, this leads to a 3-level mixed effects models with vignette judgements on level 1, individual characteristics on level 2, and vignette properties on level 3 (Rabe-Hesketh & Skrondal, 2012 a,b). However, the low number of cases makes it impossible to estimate such models.[3] As an alternative, we estimate (linear) random intercept models (Rabe-Hesketh & Skrondal, 2012a) that account for both structures separately and compare the results. As a robustness test, we also capture the structure by estimating cluster robust standard errors (Cameron & Trivedi, 2010).

To compare the results of the linear regression models used to analyze the vignette models on behavioral intentions and the logistic regression models used to analyze actual behavior, we estimate average marginal effects (AME) (Mood, 2010) for the logistic regression models. As such, they are comparable to effects estimated in linear regression models. The results of all models then display the impact in percent on the likelihood of re-entering employment either as behavioral intention or as actual behavior. To test the difference between the models obtained, we rely on two different strategies: on the one hand, we adopt a strategy proposed by Auspurg and Hinz (2011) and test whether the squared differences of regression coefficient and AME normed by the sum of both variances differ from zero. The distribution of the value of the test statistics follows a Chi-square distribution. What

[3] Due to the low number of cases, the likelihood estimators in those models do not converge.

is problematic is that this test requires no covariance between both models – an assumption that is violated per se when estimating effects for the same group under study. On the other hand, we apply a seemingly unrelated regression (SUR) (Zellner, 1962; Cameron & Trivedi, 2010). However, this strategy is unable to capture the nested structure of the data completely and can only be applied to clustered data. Thus, we can never fully capture the structure of the data.

Real-world Validation: Re-entry Intentions and Realized Re-entries

To provide a real-world validation, we examine the influence of willingness to accept unfavorable job characteristics and its impact on realized re-entries. The vignette characteristics and the individual variables are identical to the variables used in the factorial survey. The results of different specifications of the model are displayed in Table 2.

We estimate four different model specifications of logistic regression models[4] on the likelihood of re-entering the labor market. Model 1 is a random intercept model without personality traits. Central to our model is the impact of the individual deviance. Indeed, the individual deviance displaying the re-entry intention of one person as compared to somebody else confronted with the same vignette has a positive impact on the likelihood of actually re-entering the labor market in reality. The effect is significant at the 0.05 level. Thus, the higher the willingness to pay for favorable job characteristics in general, the higher also is the likelihood that somebody re-enters the labor market. At first glance, it seems counterintuitive why the vignette characteristics are still included in the models, but only when doing so, we approach the controlled net effect of the impact of the individual deviance. Furthermore, we assume that it captures the effect of time-stable unobserved heterogeneity arising from characteristics we cannot control for. From the vignette characteristics, only working fewer hours than planned as compared to more than planned increases the likelihood of re-entering the labor market.

In model 2, we also included personality traits. From those, only extraversion has a significant, positive impact on re-entry. Extraverted means that individuals are engaged with the external world and enjoy company or are active, for example. Thus, it seems rather plausible that those individuals are more likely to re-enter. Also, we can see almost identical results as compared to model 2. Again, the impact of the individual deviance from the average vignette judgment is positive and significant. When we transfer the odds ratio of the variable into an AME (Mood,

[4] The results are displayed as odds ratios (ORs). A value of the OR greater than one can be interpreted as a positive influence and a value less than one as a negative influence on the dependent variable.

Table 2 Big Five and realized job entries, different model specifications

	(1) without	(2) person ri	(3) vignette ri	(4) xtmixed
vignette characteristics				
individual deviance	1.293*	1.292*	1.116	1.009*
(standardized)	(0.144)	(0.143)	(0.0764)	(0.004)
phase: just started	1.146	1.144	1.065	1.006
ref. already searching for a while	(0.210)	(0.208)	(0.113)	(0.006)
situation: no open applications left	1.142	1.143	1.073	1.005
ref. some applications left	(0.169)	(0.167)	(0.113)	(0.007)
training: slightly over-qualified	1.261	1.260	1.151	1.009
ref. clearly over-qualified	(0.262)	(0.259)	(0.141)	(0.008)
training: according to training/abilities	1.059	1.061	1.065	1.004
	(0.228)	(0.226)	(0.140)	(0.008)
working hours: as desired	1.039	1.040	1.075	1.003
ref. more than planned	(0.245)	(0.243)	(0.143)	(0.008)
working hours: less than planned	1.452$^{\#}$	1.451$^{\#}$	1.206	1.014$^{\#}$
	(0.317)	(0.313)	(0.151)	(0.008)
commuting time: 15 minutes	1.206	1.207	1.122	1.008
ref. 45 minutes	(0.235)	(0.233)	(0.143)	(0.008)
commuting time: 30 minutes	0.990	0.992	0.996	1.000
	(0.217)	(0.216)	(0.131)	(0.008)
wage: 10 percent less	1.029	1.031	1.067	1.001
ref. according to previous job	(0.240)	(0.238)	(0.137)	(0.008)
wage: 30 percent less	1.008	1.010	1.019	0.999
	(0.212)	(0.210)	(0.139)	(0.008)
working hours: flexible	1.015	1.015	1.006	1.000
ref. fixed	(0.222)	(0.220)	(0.130)	(0.008)
working hours:	0.962	0.963	1.009	1.000
agreed upon with supervisor	(0.223)	(0.221)	(0.128)	(0.008)
individual characteristics				
partner employed full-time	4.380$^{\#}$	4.275$^{\#}$	1.796**	1.043$^{\#}$
ref. partner employed less than full-time	(3.516)	(3.449)	(0.326)	(0.025)
child under 6 in household ref. no	0.963	1.142	1.058	1.002
	(0.635)	(0.806)	(0.228)	(0.030)
age (in years)	1.062	1.066	1.021	1.002
	(0.048)	(0.052)	(0.014)	(0.002)
unemployed ref. not unemployed	4.253**	4.363**	1.916***	1.050*
	(2.137)	(2.264)	(0.210)	(0.021)

Table 2 continued

	(1) without	(2) person ri	(3) vignette ri	(4) xtmixed
tertiary education ref. no	1.172 (0.527)	1.089 (0.535)	0.896 (0.110)	0.996 (0.021)
duration of interruption (in years)	0.999 (0.038)	0.999 (0.041)	1.008 (0.010)	1.001 (0.002)
living in new federal states	1.774 (1.030)	1.819 (1.130)	1.236 (0.234)	1.026 (0.028)
first cohort ref. cohort 2	1.947 (0.910)	2.049 (1.023)	1.471^{***} (0.167)	1.028 (0.020)
participation group	0.834 (0.378)	0.873 (0.410)	0.825 (0.102)	0.984 (0.020)
Big 5: extraversion		1.044^{*} (0.021)	1.016^{**} (0.005)	1.001 (0.001)
Big 5: openness		0.995 (0.016)	0.999 (0.004)	1.000 (0.001)
Big 5: neuroticism		0.988 (0.025)	$0.988^{\#}$ (0.007)	0.999 (0.001)
Big 5: conscientiousness		1.009 (0.024)	1.007 (0.006)	1.000 (0.001)
Big 5: agreeableness		0.962 (0.028)	0.978^{**} (0.008)	0.999 (0.001)
random intercept standard deviation	4.216	4.009	0.002	
(sigma_u)	0.844	0.830	0.000	
individuals	376	376	200	
observations	3725	3725	3725	3725

Exponentiated coefficients (OR); Standard errors in parentheses
$^{\#}$ p < .10, * p < .05, ** p < .01, *** p < .001

2010), we can conclude that a one-point increase in the individual (standardized) deviance increases the re-entry probability by 26 (25.62) percentage points. In both models 1 and 2, this variable was included and we find a small positive and significant effect (at the 0.05 and 0.1 significance levels). Thus, the more willing a woman is to accept unfavorable job characteristics as examined in the factorial survey, the more likely it is that she takes up a job with less favorable characteristics. The estimated intra-class correlation rho is high (0.83). In sum, intentions and realized decisions are closely related, thereby finding preliminary support for hypothesis 1.

With respect to hypothesis 2 on the impact of personality traits, we find evidence against this because personality at least partly matters.

However, in models 1 and 2, individual characteristics become more important: Having a partner who is employed full-time compared to a partner who is employed less than full-time increases the likelihood of re-entering the labor market. This result is counterintuitive but can eventually be attributed to the low variation of the variable, which can also be seen in the high values of the standard error of this variable. Additionally, being registered as unemployed increases the likelihood of re-entering the labor market, which is in line with general expectations.

Models 3 and 4 incorporate the idea that vignette judgements are also nested in vignettes. When we look at the results of model 3 where we set the random intercept on the vignette and not the person level, we find no significant influences on the vignette level any more. Furthermore, the intra-class correlation is almost zero, indicating that modeling this structure is not necessary. Model 4 includes the deck level and the individual level in the analysis but is estimated with a mixed effects model. Again, none of the vignette characteristics is highly significant. This is also the case for personality characteristics.

Impact of Big Five Items on Job Acceptance Intentions

First, we examine the impact of the Big Five items on job acceptance intentions. Table 3 displays the results of different specifications of the model.

In model 1, we estimate a standard model including vignette characteristics and control variables without including personality traits. This model serves as a control model for our other results and has been discussed extensively from both a theoretical and empirical point of view in Drasch (2013). On the vignette level, apart from the search phase, all other characteristics have a significant impact on the willingness to accept a job offer. With respect to the other vignette characteristics and control variables on the individual level, the results remain rather stable compared to substantial research on this topic. In sum, we can conclude that mothers are willing to pay for better job characteristics in the sense that they favor jobs that are assumed to be more easily reconciled with family obligations. The intra-class correlation rho corrected for the number of variables amounts to 40.8 percent, and the LR test of the random intercept model against a linear regression model is significant, indicating that incorporating the data structure in the modeling approach as a multi-level model is indeed necessary.

The second model includes the Big Five personality traits. The model shows that apart from the personality trait conscientiousness, none of the other personality traits displays a significant effect on the willingness to accept a job offer. Thus, people who display a high level of responsibility for themselves as well as for others and who are organized, hardworking and ambitious are more likely to take up a fictive job offer. The effect itself is small. For example, a 10-point increase in the value of the conscientiousness scale increases the acceptance rate by about 2.8 percent. All other personality traits do not matter. The effects are insignificant and, aside from that, almost zero. In addition, the intra-class correlation becomes smaller, indicating less explanatory power of a model with personality traits than without. In sum, we find some weak evidence against hypothesis 2 that personality traits and behavioral intentions are not related.

Model 3 shows the coefficients of a standard linear regression model with cluster robust standard errors (Cameron & Trivedi, 2010). The results of this model are

Table 3 Big Five and job acceptance intentions, random intercept and clustered models

	(1)	(2)	(3)
	standard	+ Big 5	cluster
vignette characteristics			
main phase: just started	-0.376	-0.376	-0.376
ref. already searching for a while	(0.648)	(0.647)	(0.556)
situation: no open applications left	2.118**	2.121**	2.223***
ref. some applications left	(0.649)	(0.649)	(0.657)
training: slightly over-qualified	5.417***	5.415***	5.339***
ref. clearly over-qualified	(0.801)	(0.801)	(0.888)
training: according to training/abilities	9.056***	9.059***	9.039***
	(0.798)	(0.797)	(0.876)
working hours: as desired	15.58***	15.58***	15.62***
ref. more than planned	(0.800)	(0.799)	(1.014)
working hours: less than planned	8.806***	8.812***	8.850***
	(0.794)	(0.793)	(0.987)
commuting time: 15 minutes	22.41***	22.41***	22.27***
ref. 45 minutes	(0.794)	(0.793)	(1.124)
commuting time: 30 minutes	15.07***	15.06***	14.91***
	(0.798)	(0.797)	(0.991)
wage: 10 percent less	-5.102***	-5.091***	-5.058***
ref. according to previous job	(0.803)	(0.803)	(0.840)
wage: 30 percent less	-18.40***	-18.39***	-18.21***
	(0.797)	(0.796)	(1.038)
working hours: flexible	8.757***	8.754***	8.857***
ref. fixed	(0.797)	(0.797)	(0.937)
working hours: agreed upon with supervisor	7.084***	7.077***	7.297***
	(0.797)	(0.797)	(0.911)

Table 3 continued

	(1) standard	(2) + Big 5	(3) cluster
individual characteristics			
partner employed full-time	0.786	1.095	0.0399
ref. partner employed less than full-time	(2.131)	(2.113)	(2.248)
child under 6 in household ref. no	0.823	1.041	1.958
	(2.693)	(2.661)	(2.798)
age (in years)	-0.0158	0.00881	-0.0774
	(0.193)	(0.193)	(0.192)
unemployed ref. not unemployed	4.857**	4.759**	5.564**
	(1.828)	(1.807)	(1.848)
tertiary education ref. no	4.181*	4.435*	4.350*
	(1.887)	(1.872)	(1.861)
duration of interruption (in years)	0.137	0.114	0.126
	(0.180)	(0.178)	(0.195)
living in new federal states	2.412	2.603	3.314
	(2.426)	(2.402)	(2.438)
first cohort ref. cohort 2	3.476	3.540*	2.215
	(1.810)	(1.779)	(1.773)
participation group	-1.272	-1.282	-0.949
	(1.834)	(1.804)	(1.768)
Big 5: extraversion		-0.0384	-0.0200
		(0.076)	(0.079)
Big 5: openness		-0.0391	-0.0300
		(0.061)	(0.062)
Big 5: neuroticism		0.0924	0.0592
		(0.083)	(0.087)
Big 5: conscientiousness		0.284**	0.271**
		(0.091)	(0.091)
Big 5: agreeableness		0.108	0.0946
		(0.101)	(0.106)
constant	31.04***	9.273	14.90
	(7.623)	(10.74)	(10.41)
random intercept standard deviation	16.38***	16.03***	
(sigma_u)	(0.691)	(0.680)	
level 1 residual standard deviation	19.73***	19.73***	
(sigma_e)	(0.241)	(0.241)	
rho	0.408	0.398	0.267
individuals	376	376	(Pseudo-R^2)
observations	3725	3725	3725

Standard errors in parentheses; * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

almost identical to the results of the multilevel models, indicating stability of the results with respect to different estimation strategies.

Impact of Big Five Items and Realized Re-entries

Table 4 provides a different approach to answering the question of whether behavior and behavioral intentions are related. We now focus on the influence of individual characteristics on the re-entry probability on the individual level with and without controlling for personality traits. Because coefficients as provided by linear regression models and odds ratios provided by logistic regressions cannot be compared over different models (Mood, 2010; Auspurg & Hinz, 2011), we estimate AMEs for the logistic regressions models. Those are comparable over different model specifications, cohorts and samples.

Model 1 includes individual control variables. The only significant influence factors on realized re-entry for the group under study are the unemployment status and having a tertiary degree. Being registered as unemployed increases the re-entry probability by 8.3 percentage points. When we compare this to the results as provided by model 1 in Table 3 where the impact on re-entry intention was 4.8 and test for differences between those two coefficients with seemingly unrelated regression, the difference is significant. Furthermore, having a tertiary degree increases the re-entry probability by 4.1 percentage points, and the difference is significant at the 0.1 level. In addition, a seemingly unrelated regression indicates a significant difference at the 0.1 level. The Chi-2-test, however, displays no significant differences between the coefficients of both models.

Model 2 in Table 4 also includes personality traits. Again, an unemployment effect can be found: Registered unemployed women have an approximately 8.03 percent higher probability of actually re-entering the labor market, which is different from the intentions of model 2 in Table 3 where the effect amounts to 4.7 percent with significance at the 0.01 level. Testing whether both coefficients differ from each other, we find that the effect is significant at the 1 percent level. Additionally, the tertiary education effects are different. The overall model test of the seemingly unrelated regression is significant and allows for the following conclusion: including personality traits in the modeling approach seems to enlarge the differences between behavioral intentions and actual behavior.

Personality traits are not related to realized entries, none of the personality traits displays a significant influence on realized re-entry. Testing for differences compared to the re-entry intention model, it can be concluded that the effect of extraversion (0.171 vs. -0.0384 in the intentions model 2 Table 1) as well as conscientiousness (-0.069 vs. 0.284 and significant as displayed by the Chi2-Test at the 0.01 level) is different in models examining intentions versus realized re-entries. Thus, behavioral traits only have a minor impact in the intention models and no

Table 4 Big Five and realized re-entries, logistic regression model AMEs

	(1)			(2)		
	controls	SUR	Chi-2-	+ Big 5	SUR	Chi-2
individual characteristics						
partner employed full-time ref. partner employed less than full-time	4.394 (3.884)			3.753 (3.903)		
child under 6 in household ref. no	0.504 (4.395)			0.302 (4.44)		
age (in years)	0.109 (0.314)			0.136 (0.32)		
unemployed ref. not unemployed	8.327^{**} (3.169)	**		8.033^{*} (3.136)	***	
tertiary education ref. no	4.098 (3.022)	#		3.497 (3.052)	*	
duration of interruption (in years)	-0.107 (0.293)			-0.100 (0.292)		
living in new federal states	-1.658 (4.075)			-1.170 (4.108)		
first cohort ref. cohort 2	5.335 (3.265)			$5.475^{\#}$ (3.263)		
participation group	-0.977 (2.972)			-0.836 (2.992)		
Big 5: extraversion				0.171 (0.126)		***
Big 5: openness				-0.137 (0.099)		
Big 5: neuroticism				0.033 (0.137)		
Big 5: conscientiousness				-0.069 (0.149)	*	*
Big 5: agreeableness				-0.150 (0.169)		
Observations	378			378	** (Overall model)	
Pseudo R-squared	0.0590			0.0724		
Prob > chi2	0.131			0.262		
LR chi2	13.77			16.89		

Standard errors in parentheses
# p<0.10, < 0.05, ** p < 0.01, *** p < 0.001

impact in the behavioral model. In sum, this indicates that personality traits only have a minor importance, finding at least some support for hypothesis 2. Moreover, the effect sizes of the individual variables remain stable in models 1 and 2 indicating that the effects are similar not controlling and controlling for personality traits.

Summary

This paper examines the relation of behavioral intentions as measured with the FS approach and actual behavior. Furthermore, it examines the role of personality traits in shaping this relation. This is necessary because results from FS are often equated with actual behavior while in reality, intentions are measured. Furthermore, one can argue that personality traits might influence intentions and actual behavior differently because individuals with different personality traits might react differently to the fictive stimuli provided by FS. This can be examined with research that is able to examine both the impact of personality traits on behavioral intentions and actual behavior in the same context. The FS embedded in the evaluation project 'Perspektive Wiedereinstieg', which examined women who have been out of the labor force for several years but are in the process of re-entering the labor market, provided this rare opportunity.

In the real-world validation, by looking at the association between re-entry intentions and realized entries, one finds that they are significantly correlated. All in all, this points to the high external validity of vignette measurements. However, when testing for differences in individual characteristics, some differences can be found. These results are similar to those from Nisic and Auspurg (2009) with respect to differences in magnitude but mostly not in the decision of whether they have a significant influence.

In line with psychological theory, personality traits as measured by the Big Five do not (really) matter for behavior and behavioral intentions. We find a significant influence of conscientiousness on the willingness to accept a job offer, but we think that these results should not be over interpreted because the effect vanishes when looking at realized re-entries. Although it seems reasonable to assume that individuals that are more conscientious are more willing to accept less favorable jobs, we see these findings congruent with previous researchers' results which find an impact of other personality traits on decisions connected with mothers' employment. Although personality traits seem to matter in each of the studies, controlling or not controlling for them does not substantively alter the overall findings of the research work. In sum, this refers to the conclusion that it is not necessary to include the measurement of personality traits in factorial surveys, although they do have some impact.

However, there are some shortcomings associated with this study. First, the theoretical model is based on the assumption that the decision is under the volitional control of the individual. For obvious reasons, this might not be the case for labor market re-entries because, first, an employer must be found who hires a woman who often has been out of employment often for several years before an actual re-entry can be realized. To relax this severe deficit, one can argue that our study does not distinguish re-entries according to the volume of work but merely looks at a yes/no re-entry decision.

Second, more research is needed to examine methodological issues with respect to the relevance of job characteristics for job acceptance. Admittedly, to a certain extent our dimensions have been chosen arbitrarily through common sense. As such, however, they must be seen as examples of job characteristics that can be relevant. Future research could also make use of adaptive vignette designs by using information on previous jobs (e.g., Abraham et al., 2013), for example, by including previous wages and job conditions in the study. This was not possible in our survey due to data protection issues.

Furthermore, the results are naturally limited to the specific group under study, which is long-term non-employed mothers seeking employment. Future research could benefit from studying behavioral intentions, for example, job acceptance intentions from a much more diverse group. This requires a representative sample and a larger study context.

References

Ajzen, I. (1988). *Attitudes, personality, and behavior.* Chicago: Dorsey Press.

Ajzen, I. (1991). Theories of cognitive self-regulation. The theory of planned behavior. *Organizational Behavior and Human Decision Processes*, 50(2), 179-211. doi:10.1016/0749-5978(91)90020-T

Almlund, M., Duckworth, A.L., Heckman J.J., & Kautz, T.D. (2011). Personality psychology and economics. *NBER Working Paper Working Paper No. 16822,* Cambridge, Massachusetts.

Anger, S., Kvasnicka, M., & Siedler, T. (2011). One last puff? Public smoking bans and smoking behavior. *Journal of Health Economics*, 30(3), 591-601. doi:10.1016/j.jhealeco.2011.03.003

Auspurg K., & Hinz, T. (2011). Gruppenvergleiche bei Regressionen mit binären abhängigen Variablen – Probleme und Fehleinschätzungen am Beispiel von Bildungschancen im Kohortenverlauf. *Zeitschrift für Soziologie*, 40(1), 62-73.

Auspurg, K., & Hinz, T. (2015). *Factorial survey experiments. Quantitative applications in the social sciences 175.* Thousand Oaks, CA: Sage.

Auspurg, K., Hinz, T., Liebig, S., & Sauer, C. (2015). The factorial survey as a method for measuring sensitive issues. In U. Engel, B. Jann, P. Lynn, A. Scherpenzeel, & P. Sturgis (Eds.), *Improving Survey Methods. Lessons from recent research* (pp. 137-149). New York and London: Routledge.

Abraham, M., Auspurg, K., Bähr, S., Frodermann, C., Gundert, S., & Hinz, T. (2013). Unemployment and willingness to accept job offers: Initial results of a factorial survey approach. *Journal of Labour Market Research*, 46(4), 283-305. doi:10.1007/s12651-013-0142-1

Back, M.D., Schmukle, S.C., & Egloff, B. (2009). Predicting actual behavior from the explicit and implicit self-concept of personality. *Journal of Personality and Social Psychology*, 97(3), 533-548. doi: 10.1037/a0016229

Berger, E. M. (2010). Women's non-cognitive skills and transition to employment after childbirth. Germany: German Institute for Economic Research (DIW Berlin) and Freie Universitat Berlin. Retrieved March 21, 2016, from DIW website: https://www.diw.de/documents/dokumentenarchiv/17/diw_01.c.359339.de/berger_eea_2010.pdf

Cameron, A. C., & Trivedi, P.K. (2010). *Microeconometrics using Stata*. Revised Edition 2010. College Station, TX: Stata Press.

Connor, M., & Abraham, C. (2001). Conscientiousness and the theory of planned behavior: toward a more complete model of antecedents of intentions and behavior. *Personality and Social Psychology Bulletin*, 27(11), 1547-1561. doi: 10.1177/01461672012711014

Dehne, M., & Schupp, J. (2007). Persönlichkeitsmerkmale im Sozio-oekonomischen Panel (SOEP) Konzepte, Umsetzung und empirische Eigenschaften. *DIW Research Notes 26*, Berlin.

Diener, K., Götz, S., Schreyer, F., Stephan, G. (2013). *Beruflicher Wiedereinstieg von Frauen nach familienbedingter Erwerbsunterbrechung: Befunde der Evaluation des ESF-Programms „Perspektive Wiedereinstieg" des Bundesministeriums für Familie, Senioren, Frauen und Jugend IAB-Forschungsbericht (IAB Research Report), 09/2013*, Nürnberg, 109 S.

Drasch, K. (2013). *The re-entry of mothers in Germany into employment after family-related interruptions. Empirical evidence and methodological aspects from a life course perspective* (Doctoral dissertation). IAB-Bibliothek, Band 343. W. Bielefeld: Bertelsmann Verlag.

Dülmer, H. (2007). Experimental plans in factorial surveys: random or quota design?. *Sociological Methods & Research*, 35(3), 382-409. doi:10.1177/0049124106292367

Dülmer, H. (2016). The factorial survey: Design Selection and its Impact on Reliability and Internal Validity. *Sociological Methods & Research*, 45(2), 304-347. doi:10.1177/0049124115582269

Gerlitz, J.-Y., & Schupp, J. (2005). Zur Erhebung der Big-Five-basierten Persönlichkeitsmerkmale im SOEP Dokumentation der Instrumentenentwicklung BFI-S auf Basis des SOEP-Pretests 2005. *DIW Research Notes 4*, Berlin.

Hainmueller, J., Hangartner, D., & Yamamoto, T. (2015). Validating vignette and conjoint survey experiments against real-world behavior. *Proceedings of the National Academy of Sciences of the United States of America*, 112(8), 2395-2400. doi: 10.1073/pnas.1416587112

Jasso, G. (2006). Factorial survey methods for studying beliefs and judgments. *Sociological Methods & Research*, 34(3), 334-423. doi:10.1177/0049124105283121

Kalter, F. (1997). *Wohnortwechsel in Deutschland. Ein Beitrag zur Migrationssoziologie und zur empirischen Anwendung von Rational-Choice-Modellen.* Opladen: Leske+Budrich.

Heineck, G., & Anger, S. (2010). The returns to cognitive abilities and personality traits in Germany. *Labour Economics*, 17(3), 535-546. doi:10.1016/j.labeco.2009.06.001

Kuhfeld, W. F., Randall, D. T., & Garratt, M. (1994). Efficient experimental design with marketing research applications. *Journal of Marketing Research*, 31(4), 545-557.

Mood, C. (2010). Logistic regression: Why we cannot do what we think we can do and what we can do about it. *European Sociological Review*, 26(1), 67-82. doi:10.1093/esr/jcp006

Mischel. W. (1968). *Personality and assessment.* New York: Wiley.

Nisic, N., & Auspurg, K. (2009). Faktorieller Survey und klassische Bevölkerungsumfrage im Vergleich–Validität, Grenzen und Möglichkeiten beider Ansätze. In P. Kriwy & C. Gross (Eds.), *Klein aber fein! Quantitative empirische Sozialforschung mit kleinen Fallzahlen* (pp. 211-246). Wiesbaden: VS Verlag. doi:10.1007/978-3-531-91380-3_9

Rabe-Hesketh, S., & Skrondal, A. (2012a). *Multilevel and longitudinal modeling Using Stata. Volume I: Continuous Responses. Third Edition.* College Station, TX: Stata Press.

Rabe-Hesketh, S., & Skrondal, A. (2012b). *Multilevel and Longitudinal modeling Using Stata. Volume II: Categorical Responses, Counts, and Survival. Third Edition.* College Station, TX: Stata Press.

Rossi, P. H., &. Anderson, A. B (1982) The factorial survey approach. An introduction. In P. H. Rossi, & Nock S. L. (Eds.), *Measuring social judgments. The factorial survey approach* (pp.15-67). Beverly Hills: Sage.

Schaeffer, N.C., & Bradburn, N. M. (1989). Respondent behavior in magnitude estimation. *Journal of the American Statistical Association*, 84(406), 402-413. doi:10.1080/01621459.1989.10478784

Wallander, L. (2009). 25 years of factorial surveys in sociology: A review. *Social Science Research*, 38(3), 505-520. doi:10.1016/j.ssresearch.2009.03.004

Wichert, L., & Pohlmeier, W. (2010). Female labor force participation and the Big Five. *Zentrum für Europäische Wirtschaftsforschung Discussion-Paper 10-003*, Mannheim.

Zellner, A. (1962). An efficient method of estimating seemingly unrelated regressions and tests for aggregation bias. *Journal of the American Statistical Association,* 57(298), 348-368. doi:10.2307/2281644

Table A Descriptive results

Variable	N	Mean	Std. Dev.	Min	Max
vignette judgement	3,725	61.72	29.75	1.00	100.00
individual deviance vignette	3,725	0.077	26.68	-80	76.1
Big 5: extraversion	3,725	50.58	16.34	-11.19	88.79
Big 5: openness	3,725	50.57	18.59	-1.53	87.62
Big 5: neuroticism	3,725	50.23	11.01	17.28	80.90
Big 5: conscientiousness	3,725	50.71	11.55	2.19	84.11
Big 5: agreeableness	3,725	50.57	9.85	8.54	80.89
partner employed fulltime	3,725	0.77	0.42	0.00	1.00
child under 6 in household	3,725	0.17	0.38	0.00	1.00
age (in years)	3,725	42.16	6.36	25	60
registered unemployed	3,725	0.44	0.50	0.00	1.00
tertiary education	3,725	0.42	0.49	0.00	1.00
duration of interruption	3,725	10.65	6.58	0.00	29.67
living in new federal states	3,725	0.19	0.39	0.00	1.00
first cohort	3,725	0.58	0.49	0.00	1.00
participant group	3,725	0.49	0.50	0.00	1.00

methods, data, analyses | Vol. 13(2), 2019, pp. 291-306 DOI: 10.12758/mda.2019.04

Recruiting Young and Urban Groups into a Probability-Based Online Panel by Promoting Smartphone Use

Peter Lugtig[1], *Vera Toepoel*[1], *Marieke Haan*[1], *Robbert Zandvliet*[2] & *Laurens Klein Kranenburg*[2]

[1] *Department of Methodology and Statistics, Utrecht University*

[2] *I&O Research*

Abstract

A sizable minority of all web surveys are nowadays completed on smartphones. People who choose a smartphone for Internet-related tasks are different from people who mainly use a PC or tablet. Smartphone use is particularly high among the young and urban. We have to make web surveys attractive for smartphone completion in order not to lose these groups of smartphone users. In this paper we study how to encourage people to complete surveys on smartphones in order to attract hard-to-reach subgroups of the population. We experimentally test new features of a survey-friendly design: we test two versions of an invitation letter to a survey, a new questionnaire lay-out, and autoforwarding. The goal of the experiment is to evaluate whether the new survey design attracts more smartphone users, leads to a better survey experience on smartphones and results in more respondents signing up to become a member of a probability-based online panel. Our results show that the invitation letter that emphasizes the possibility for smartphone completion does not yield a higher response rate than the control condition, nor do we find differences in the socio-demographic background of respondents. We do find that slightly more respondents choose a smartphone for survey completion. The changes in the layout of the questionnaire do lead to a change in survey experience on the smartphone. Smartphone respondents need 20% less time to complete the survey when the questionnaire includes autoforwarding. However, we do not find that respondents evaluate the survey better, nor are they more likely to become a member of the panel when asked at the end of the survey. We conclude with a discussion of autoforwarding in web surveys and methods to attract smartphone users to web surveys.

Keywords: mobile surveys, autoforward, survey design, probability-based online panel

Smartphone users are different from people who mainly use a PC or laptop to access the Internet (Busse & Fuchs 2012; Couper et al. 2017; Maslovskaya et al. 2017). Over time there has been a persistent difference in correlates of coverage of smartphone users in both Europe and the United States. People who use the smartphone for Internet browsing are younger and live in more urban areas than people who use PCs, laptops or tablets (Busse & Fuchs 2012, 2014). There is a small but growing group of people who are "mobile-only" (Lugtig et al. 2016; Maslovskaya et al. 2017), which to some degree overlap with hard-to-reach respondents in general (Mac Ginty & Firchow 2017). Young and urban respondents are generally hard-to-reach in surveys; smartphone penetration and usage is also highest in this group (Haan et al. 2014).

In 2017, 60% of Dutch adults report to own a PC, 82% a laptop, 72% a tablet and 89% a mobile or smartphone. Only 39% of respondents report to have used a laptop in the last 3 months to access the Internet, 36% a tablet, and 79% a smartphone (Statistics Netherlands 2017). Despite the fact that smartphones are used often by many people, many respondents still prefer PCs or laptops to participate in surveys. Between 10-30% of all web surveys are started on smartphones (Bosnjak et al. 2018; Brosnan et al. 2017; Masvlovskaya et al. 2017). This "gap" between the frequent use of smartphones in general, and infrequent use of them in web surveys can probably be explained by respondents' expectations and experiences of completing web surveys on smartphones.

Two differences stand out when the web survey experience on PCs and smartphones are compared. First, the screen on smartphones is much smaller, leading to challenges in presenting complex survey questions. Without adaptations, web survey question texts may not fit a smartphone screen, forcing respondents to scroll vertically or horizontally. Several studies have shown that splitting up grids and displaying one or a few items per page when participating on a mobile phone is a good solution to this problem (Keusch & Yan 2016; Mavletova & Couper 2016; Antoun et al. 2017). Still, even in this format, and controlling for respondent and question characteristics, Couper & Peterson (2017) find that mobile surveys take longer to complete. Mavletova & Couper (2015) moreover find that break-off rates are generally higher when respondents complete a web-survey on a smartphone, and that breakoff rates are considerably higher when web surveys are not optimized for smartphones. Despite web surveys becoming smartphone-completable, they are often still not smartphone-friendly or smartphone-optimized (Revilla et al. 2017). Designing surveys to be smartphone-friendly is necessary to convince potential

Direct correspondence to
Peter Lugtig , Department of Methodology and Statistics, Utrecht University,
Padualaan 14, 3508 TC Utrecht, the Netherlands
E-mail: p.lugtig@uu.nl

respondents that surveys can be completed on smartphones, and to make sure that they do not drop out.

The second difference has to do with how respondents navigate from question to question, and page to page. PC respondents use a mouse and keyboard whereas smartphone respondents use their fingers. Answer selection is often done with radio buttons, while page-to-page navigation is typically done with ‘next’ and ‘back’ buttons. Some smartphone respondents may have trouble selecting those answers, or even finding them, especially when the ‘next’ and ‘back’ buttons are at the bottom corners of the page.

With autoforwarding respondents no longer have to press the ’next’ button to move to the next page. Instead they auto-advance, auto-submit or auto-forward to the next question after an answer is given. An early study on autoforwarding (Hays et al. 2010) focusing on PC users showed that autoforwarding may shorten the completion time, but may come at the expense of losing a smooth navigation experience. Respondents using the PC may be familiar with clicking multiple times to advance from page-to-page. More recent studies investigated autoforwarding specifically in the context of the rise of smartphones in web surveys (Arn et al. 2015). They have found that autoforwarding may work well with easy questions (Selkälä & Couper 2017), and can shorten response times (de Bruijne 2016), although there is a risk that respondents may find autoforwarding confusing, or need longer to think about an answer, especially when the questionnaire consists of cognitively difficult questions.

This study reports on an experimental survey that had the twin goal of convincing potential respondents to start a survey on their smartphone, and to deliver a better survey experience by a better layout and eliminating any need to scroll on the smartphone. We tested two versions of an invitation letter, in which one version emphasized that the survey was smartphone-friendly, while the other did not. These two conditions were crossed with 3 different versions of the questionnaire: 1) the old, not-smartphone optimized layout, 2) a new, smartphone-friendly layout of the questionnaire, and 3) a smartphone friendly layout combined with autoforwarding. We expect that the new invitation letter leads to a higher proportion of people who start the survey on a mobile phone, and also that those respondents are younger and come from more urban areas. We expect that the new questionnaire layout leads to shorter completion times, a better survey experience, and ultimately, more respondents who sign up to become a panel member.

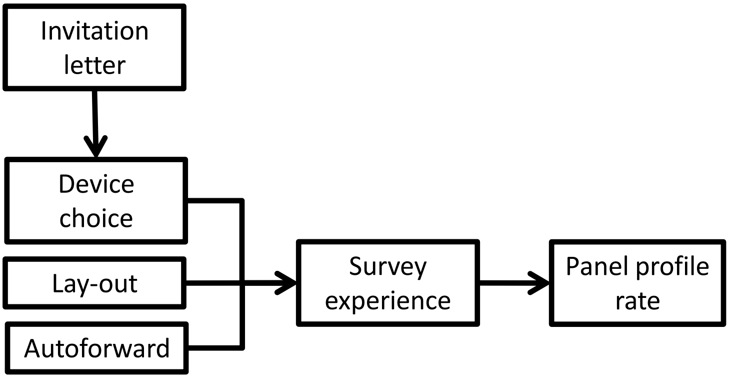


Figure 1 Theoretical model. The type of advance letter (mobile phone emphasis or not) determines device choice. Device, along with the experimental layout and navigation conditions determines the survey experience, which in turn affects how likely respondents become a panel member

Methods

Sample and Recruitment

The goal of the survey was to recruit respondents for the I&O Research Panel. This is a probability-based online panel of people in the Netherlands. Panelists are recruited through an opt-in question asked at the end of a recruitment survey that is fielded twice a year.[1] In this paper, we use data from the second round of 2017, which was fielded on August 30, 2017. A random sample of addresses selected from the postal address file of the Netherlands was invited by mail to participate in an online survey about trends in Dutch society. Because panelists from particular regions and urban areas in the Netherlands were underrepresented in the I&O Research Panel, the sample was a two-stage stratified cluster sample, with clusters consisting of 20 cities with a population larger than 100,000 inhabitants and 5 provinces (Friesland, Groningen, Flevoland, Noord-Brabant and Limburg). In each of the 20 cities[2], 1500 addresses were selected using simple random sampling, while in each of the provinces (excluding cities already selected within those provinces), 6000 addresses were selected. The survey stayed online until October 1st. Respondents could always leave the survey, and start where they left off at a later

[1] In order to become a member of the I&O Research Panel, people have to complete a double opt-in procedure.

[2] The cities included are Amsterdam, Rotterdam, Tilburg, Breda, Eindhoven, Nijmegen, ‘s Hertogenbosch, Arnhem, Dordrecht, Ede, Apeldoorn, Zwolle, Oss, Maastricht, Roosendaal, Bergen op Zoom, Hilversum, Sittard-Geleen, Doetinchem and Heerlen.

moment. A raffle was held among all participants, in which five 100-euro and twentyfive 20-euro gift vouchers for a popular online-shopping website were given. No reminders were used. In our study, we used unweighted data[3].

Experiment with the Invitation Letter

The invitation letter to the survey asked whomever opened the letter to give the letter to the youngest person living in the household over the age of 16. The reason for this is again the fact that young people were underrepresented in the I&O panel. The invitation letter mentioned that the survey was about safety, social media, health and leisure, and provided a URL with individualized login to the survey. Within the invitation letter we embedded an experiment: in the old version of the letter (conditions 1, 2 and 4) we showed an icon of a regular PC next to the URL. In another version (conditions 3 and 5), we replaced the icon of the PC with a mobile phone, and included an additional sentence below the URL that stated that the survey was easy to complete on PCs, tablets or mobile phones. The goal of this experiment was to test whether 1) respondents would be more likely to use a mobile phone for completion and 2) whether we could attract young and urban respondents at a higher rate. The two versions of the invitation letter are shown in Appendix A.

Experiment Within the Questionnaire

Within the questionnaire, we experimented with the layout and navigation, split into three conditions, shown visually in Appendix B. In all three conditions, questions were presented page-by-page, and all versions used radio buttons. The versions differed however in the following ways:

1. A condition in which the old layout was used, not optimized for smartphones (condition 1).
2. A condition in which the new layout optimized for smartphones was used. In this layout answer options were presented vertically, and the width of the questionnaire was automatically adapted to the size of the screen (conditions 2 and 3).
3. A condition that was identical to the new layout used in conditions 2 and 3, with autoforwarding added to this (conditions 4 and 5). When a respondent selected an answer, the next question was automatically shown on a new page. A ‘forward’ and ‘back’ button were still present so that respondents could skip a question or correct an earlier answer. Autoforwarding was used throughout the entire questionnaire.

[3] We repeated our analyses using sampling weights which correct for unequal selection probabilities of households across strata and clusters in our sample, but found no meaningful differences in the results.

Table 1 Experimental design of study

	Invitation letter	Smartphone friendly layout	Autoforwarding	Gross sample size
Condition 1	Old	No	No	12000
Condition 2	Old	Yes	No	12000
Condition 3	New	Yes	No	12000
Condition 4	Old	Yes	Yes	12000
Condition 5	New	Yes	Yes	12000

Notes: The new letter included an icon of a smartphone, as well as the note that the survey could be completed on all devices. The old letter showed an icon of a PC.

Table 1 summarizes the design of our study. In total, we used 5 conditions, in which elements from the invitation and questionnaire experiments were combined. At the end of the recruitment survey, respondents received the question whether they would like to become a member of the I&O Research Panel. The new smartphone-friendly layout was responsive to smartphones. The old-layout was not responsive.

Analysis

We study whether the invitation experiment leads to a different response rate across devices and a different composition of the respondents. Then we study whether the layout and autoforwarding experiments lead to shorter survey completion times, a better evaluation of the survey and a higher proportion of respondents becoming a panel member (Profile Rate or PROR (Callegaro & DiSogra 2008)).

In order to determine what device respondents used to complete the survey, we coded every device that was used at the start of the survey. Devices with a screensize of 6.0 inches or smaller were defined as ‘smartphone’. Devices with a screensize larger than this were defined as PC/tablet. Because we compare multiple groups on different variables, we choose to conduct significant tests with $\alpha = .005$ (Benjamin et al. 2018).

Results

Response to the Survey Across Invitation Letter Conditions

Table 2 shows the effect of the invitation letter on response rates and response composition in the recruitment survey. We find that the response rate for the recruitment

Table 2 Composition of response in recruitment survey

	Old letter	New letter	Statistical difference test
Response rate	6.01%	6.22%	$\chi^2(1)=1.09$, $p=.29$
Smartphone completion within responses	19.0%	23.1%	$\chi^2(1)=9.36$ $p<.005$
Within smartphone respondents			
Young (<25 year)	26.6%	25.2%	$\chi^2(1)=.18$, $p=.67$
From a big city (> 100.000 inhabitants)	52.4%	50.7%	$\chi^2(1)=.22$, $p=.64$
Within PC respondents			
Young (<25 year)	13.1%	14.1%	$\chi^2(1)=.60$, $p=.44$
From a big city (> 100.000 inhabitants)	48.4%	47.8%	$\chi^2(1)=.12$, $p=.73$

interview using the old letter (conditions 1, 2 and 4) is 6.01%, and for the new letter (conditions 3 and 5) is 6.22%. This difference is not significant.

Although the invitation letter does not lead to a higher response rate, we do see a difference in the proportion of respondents using a smartphone. When respondents receive the old letter 19.0% decide to use a smartphone, whereas this is 23.1% when they receive the new letter (see Table 2). In terms of the composition of the response, we find that the new letter does not lead to younger people, or people living in urban areas being more likely to respond.

Effects of Device, Layout and Autoforwarding on Survey Experience

Respondents can choose themselves what device they use for survey completion. As a consequence, there will be self-selection effects between PC and smartphone respondents when we study the survey experience. To account for the selection effects, we split the following analyses by respondents who completed the survey on a PC and a smartphone.

First, we look at survey completion times. The new layout and autoforwarding should lead to a relatively shorter completion time on smartphones. For PC respondents, Table 3 shows that median completion times in the old design (condition 1=10.1), the new design (condition 2 and 3 = 10.5) and the new design + autoforward do not differ (conditions 4 and 5 = 10.5). When respondents complete the survey on smartphones, the completion time is shorter in the new design, and when autoforwarding is used (medians in conditions 1, 2-3 and 4-5 = 10.5, 9.9 and 9.5, Kruskal-Wallis Test *p*-value <.005). We conclude that response times were about

Table 3 Net response rate, completion time, survey evaluation and profile rates split for device used, across 5 experimental conditions.

	Device used	Net response recruitment survey	Survey completion time in minutes (median)	Mean survey evaluation (standard deviation)	Panel members	Panel profile rate
Condition 1	PC	612	10.1	7.6 (1.2)	419	68.5
	Smartphone	135	10.5	7.4 (1.4)	85	63.0
Condition 2	PC	568	10.4	7.5 (1.3)	382	67.3
	Smartphone	137	10.0	7.5 (1.3)	96	70.1
Condition 3	PC	585	10.6	7.4 (1.4)	405	69.2
	Smartphone	152	9.9	7.5 (1.3)	104	68.4
Condition 4	PC	571	10.2	7.5 (1.2)	390	68.3
	Smartphone	138	9.0	7.7 (1.4)	90	65.2
Condition 5	PC	562	10.1	7.5 (1.3)	386	68.7
	Smartphone	193	10.0	7.7 (1.3)	136	70.5

Notes: See Table 1 for explanation of the experimental conditions. The panel profile rate is calculated conditional on respondents starting the recruitment survey. The total response rate of the survey is 6.23% in condition 1, 5.89% in condition 2, 6.14% in condition 3, 5.91% in condition 4, and 6.29% in condition 5. The unconditional panel recruitment rate is 4.20% in condition 1, 3.98% in condition 2, 4.24% in condition 3, 4.00% in condition 4, and 4.35% in condition 5.

the same in all conditions when respondents used a PC, but about 20% shorter when the new layout and autoforwarding were used for smartphone respondents.

At the end of the recruitment survey, respondents were asked to rate the survey experience on a scale from 1 (very bad) to 10 (very good). Only the endpoints were labelled. Smartphone respondents on average evaluate the questionnaire with a 7.4, 7.5 and 7.7 in conditions 1, 2-3 and 4-5. This difference among smartphone respondents is not significant (F(2,728)=2.97, p=.05). This implies that despite the shorter time it took to complete the survey, smartphone respondents were not happier with the new layout and autoforwarding.

Effects on Panel Membership

Finally, we study whether the combined effect of the new letter and the questionnaire experiments have any effect on the panel membership rate. When we look at the PC respondents only, we find no differences between conditions. The panel profile rate, conditional on starting the survey, ranges from a low of 67.3% in condition 3 to a high of 68.7% in condition 5 for PC respondents. For smartphone respondents, we find no effect for panel membership either. The profile rate in conditions with the old layout is 63.0%, 69.2% in the new layout and 67.8% in the new layout combined with autoforward. A test across the three layout conditions showed that this difference is not significant ($\chi^2(2)$=1,73, *p*=.42). There is a strong relationship between the survey evaluation and panel membership however. People who did not become a panel member give the survey a 6.9 on average, whereas panel members give the survey a 7.8 on average. In terms of the theoretical model shown in Figure 1, we have to conclude that the survey experience does affect the panel membership rate. Our experimental manipulations does however not result in respondents being happier with the survey, despite the reduction in the time to complete the survey.

Response Quality

We finally take a look at response quality, as a further exploratory analysis of the effects of our experimental conditions and to understand whether the reduction in interview time on smartphones comes at the price of lower data quality. Earlier studies have indicated that respondents may sometimes inadvertently skip questions in the autoforward condition, or otherwise have trouble navigating the questionnaire. Do we find evidence for this in our data? We do not have validation data in order to check whether the data respondents provided is accurate, nor detailed audit trail data, nor did the questionnaire include response scales which allow for psychometric modeling of data quality. We therefore rely on indirect indicators of data quality, which have been used before by for example Kaminska & Lynn (2012) and Lugtig & Toepoel (2016) to model data quality of smartphone survey responses. Specifically, we look at five sets of indicators:

1. whether respondents finished the questionnaire, how many questions were not answered, and how many times respondents answered “Don’t know”.
2. two indicators for response behavior in scales: For straightlining, whether at any point in the questionnaire the respondent gives the same answer to all items on the following scales: a three-item scale asking about the difficulty of completing forms, a 12-item scale asking about the frequency of leisure activities, a 7-item scale asking about the importance of aspects of life (family, friends, leisure time, politics, work, religion, school), a 4-item scale asking about interest in food, and a 4-item scale asking about fear for terrorism. If the respondent straightlined on any of these scales, we assigned a score of 1, and if not, we assigned a 0.
3. we also code how many answers respondents choose in 2 check-all-that-apply questions asking for the use of 14 types of social media, and consumption of

11 types of new sources. More answers are considered to be indicative of better data.

4. we code the primacy effect by counting how often respondents clicked the first answer on three scales. The two scales mentioned above, and a third scale, where respondents were asked to indicate what should be the priorities for the Netherlands from a list of 24 policy-issues. The occurrence of a primacy effect is a sign of lower data quality
5. finally we check whether respondents left a comment to the final question "do you have any remarks" and if so, we count how many characters were included in these answers. Longer answers are considered better.

Table 4 shows the differences between the three different questionnaire layout conditions, split for the device that respondents used. Across the 8 indicators that we distinguish to study data quality, we only find differences for 3 of them. For the number of "don't know" responses, we find more don't know responses on smartphones and fewer don't know responses when autoforward is used. For both straightlining and the number of answers chosen in the check-all-that apply question, we find that smartphone users provide better quality: they straightline less, and provide more answers in the check-all-that apply question. As respondents could self-select their device, it is likely that the differences we find here are self-selection effects.

Most striking is that we find no other effects for any of the experimental conditions, nor any interactions between our experiment and device used. This implies that we find that the answers that respondents give to our questionnaire do not depend on the questionnaire layout. No matter what layout condition respondents get, the answers they provide are of about the same quality. Faster responses in the smartphone friendly and autoforwarding conditions do not come as a price of lower data quality.

Table 4 Response quality indicators across the three layout questionnaires, split for the device used by the respondent

	Old layout (not friendly)		Smartphone friendly layout		Friendly layout + autoforwarding	
Device	PC	Smartphone	PC	Smartphone	PC	Smartphone
Dropout %	4.4	3.7	3.3	3.9	2.7	2.5
Mean Item missing	.27	.34	.26	.33	.26	.27
Mean "Don't know" answers	.66	1.33	.68	.86	.67	.69
Any Straightlining	.34	.23	.34	.27	.34	.29
# answers chosen in 2 check-all-that-apply questions	7.22	7.45	7.23	7.81	7.21	8.13
# Primacy effect (max=3)	1.81	1.86	1.82	1.77	1.80	1.92
% Left a comment	17	21	18	18	16	17
Mean character length of comment if comment given	100	61	112	77	98	58

Notes: Univariate ANOVA Tests per behavior with layout, autoforward, device used and interactions between these variables as factors. Findings: Dropout: no effects, Item missings: no effects, DK: main effect of device, autoforward, Straightlining: effect of device, Check-all-that-apply: Effect of device, Primacy effect: no effects, Left a comment: no effects, Length of comment: no effects.

Discussion

In this study, we tested two ways to recruit smartphone-users into a probability-based online panel. We find that a new invitation letter, emphasizing the possibility to participate on smartphones, does not lead to any more or different respondents participating in an online-recruitment survey when compared to an old letter emphasizing PCs. Respondents are however somewhat more likely to use a smartphone to complete the recruitment interview. One possible cause of our null-finding is the relatively small change in the introduction letter: we used a sentence and changed an icon, but did not use specific fonts, or changed the content of the letters.

A second experiment had the aim to ensure that respondents who started the recruitment interview were more likely to become a panel member. A new responsive questionnaire layout and method of navigation had the aim to make the survey experience shorter and more enjoyable. Here we find that that the new layout and autoforwarding lead to a reduction of about 20% in the survey completion time for smartphone respondents, but that smartphone respondents do not evaluate the new questionnaire more positively. Consequently, we find no differences in the panel profile rate across conditions.

The reduction in response times does not appear to have led to lower data quality. We believe that this is a promising finding. This contrasts the findings of for example de Bruijne (2016) who found that respondents skip questions when autoforwarding is used. Perhaps this has to do with the fact that we use a fresh cross-sectional sample instead of experienced panel members. The fact that we use a cross-sectional survey also leads to a limitation: the response rates in our study were very low. Using reminders or incentives could help to increase these and perhaps alter our findings with regards to the invitation letter.

We do not believe that a higher response rate would lead to differences in our results for the questionnaire design experiments. The panel profile rate conditional on response will probably decrease when more difficult to reach respondents participate in the recruitment interview, but it is hard to imagine that harder-to-reach respondents respond differently to the questionnaire layout designs we tested.

Selkälä & Couper (2017) argue that autoforwarding mainly works for questions that require little cognitive effort. Our survey consisted of relatively simple questions asking about a variety of topics. It thus remains to be seen whether the reduction in completion times that we observe also holds in other studies. One way to understand the response behavior for different types of questions is by studying audit trails, which can be collected with web surveys, and look at response behavior in more detail. We did however not observe accidental skipping of questions in the autoforward conditions, as was reported by de Bruijne (2016).

Despite the fact that respondents who use a smartphone in our study need less time, they do not evaluate the survey to be better. Perhaps our experimental manipulation was not strong enough; we still used radio buttons for example in all conditions. Making questionnaire more smartphone friendly is still a big challenge for survey research. Smartphone users are generally younger and live in more urban areas; two of the characteristics of respondent groups who are hard-to-recruit in many countries. How should we design our surveys so that these people are more likely to participate? There are perhaps ways in which we can make the recruitment process more attractive to smartphone users. Adding a QR code, NFC-tag or other measures to facilitate the transition from a paper invitation letter to a questionnaire in a smartphone browser may help somewhat, but a large challenge remains for future survey research. In an era of declining response rates and diversifying device use, how can we design surveys so that there are as few barriers to survey completion as possible?

References

Antoun, C., Katz, J., Argueta, J., & Wang, L. (2017). Design Heuristics for Effective Smartphone Questionnaires. *Social Science Computer Review*, doi:10.1177/0894439317727072.

Arn, B., Klug, S., & Kolodziejski, J. (2015). Evaluation of an adapted design in a multi-device online panel: A DemoSCOPE case study. *methods, data, analyses*, *9*(2), 28.

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E.- J., Berk, R., ... & Cesarini, D. (2018). Redefine statistical significance. *Nature Human Behaviour*, *2*(1), 6.

Bosnjak, M., Bauer, R., & Weyandt, K. W. (2018). Mixed Devices in Online Surveys: Prevalence, Determinants, and Consequences. In Theorbald, A. (ed). *Mobile Research* (pp. 53-65). Springer Gabler, Wiesbaden.

Brosnan, K. Grün, B., & Dolnicar, S. (2017). PC, Phone or Tablet?: Use, Preference and Completion Rates for Web Surveys. *International Journal of Market Research*, *59*(1), 35-55.

Busse, B., & Fuchs, M. (2014). Recruiting Respondents for a Mobile Phone Panel. *Methodology, 10.* 21-30. Doi: 10.1027/1614-2241/a000064

Busse, B., & Fuchs, M. (2012). The components of landline telephone survey coverage bias. The relative importance of no-phone and mobile-only populations. *Quality & quantity*, *46*(4), 1209-1225.

Callegaro, M. (2010). Do you know which device your respondent has used to take your online survey? *Survey Practice*, *3*(6).

Callegaro, M., & DiSogra, C. (2008). Computing Response Metrics for Online Panels, *Public Opinion Quarterly*, *72*(5), 1008-1032, doi: 10.1093poq/nfn065

Couper, Mick P., Antoun, Christopher, & Mavletova, A. (2017). Mobile Web Surveys. In P.P. Biemer, E.D. de Leeuw, S. Eckman, B. Edwards, F. Kreuter, L.E. Lyberg, N.C. Tucker, B.T. West (Eds). *Total Survey Error in Practice*. New York: Wiley. 133-154.

Couper, M. P., & Peterson, G. J. (2017). Why do web surveys take longer on smartphones?. *Social Science Computer Review*, *35*(3), 357-377.

De Bruijne, M. (2016). Online vragenlijsten en mobiele devices (Online questionnaires and mobile devices). *Jaarboek van de Marktonderzoeksassociatie*, 137-15. Available on http://moa04.artoo.nl/clou-moaweb-images/images/bestanden/pdf/Jaarboeken_MOA/Marktonderzoek_2016_H9.pdf

Haan, M., Ongena, Y. P., & Aarts, K. (2014). Reaching hard-to-survey populations: Mode choice and mode preference. *Journal of Official Statistics*, 30(2). 355–379

Hays, R. D., Bode, R., Rothrock, N., Riley, W., Cella, D., & Gershon, R. (2010). The impact of next and back buttons on time to complete and measurement reliability in computer-based surveys. *Quality of Life Research*, *19*(8), 1181-1184.

Keusch, F., & Yan, Ting (2016). Web versus mobile web: an experimental study of device effects and self-selection effects. *Social Science Computer Review*, doi:10.1177/0894439316675566.

Lugtig, P., Toepoel, V., & Amin, A. (2016). Mobile-only web survey respondents. *Survey Practice*, *9*(4).

Lynn, P., & Kaminska, O. (2012). The impact of mobile phones on survey measurement error. *Public Opinion Quarterly*, *77*(2), 586-605.

Mac Ginty, R.,, & Firchow, P. (2017). Including Hard-to-Access Population Using Mobile Phone Surveys and Participatory Indicators. *Sociological Methods & Research*. DOI: 10.1177/0049124117729702

Maslovskaya, O., Durrant, G. Smith, P. W.F., Hanson, T., & Villar, A. (2017). Mixed-device online surveys in the UK. NCRM working paper 4/17.

Mavletova, A and Couper, M P. (2015). A Meta-Analysis of Breakoff Rates in Mobile Web Surveys. In D. Toninelli, R, Pinter, & P. de Pedraza (eds.), *Mobile Research Methods: Opportunities and Challenges of Mobile Research Methodologies* (pp. 81–98). London: Ubiquity Press. DOI: http://dx.doi.org/10.5334/bar.f.

Mavletova, A., & Couper, M. P. (2016). Grouping of items in mobile web questionnaires. *Field Methods*, *28*(2), 170-193.

Mavletova, A., Couper, M. P., & Lebedev, D. (2017). Grid and Item-by-Item Formats in PC and Mobile Web Surveys. *Social Science Computer Review*, doi: 10.1177/0894439317735307.

Revilla, M., Ochoa, C., & Turbina, A. (2017). Making use of Internet interactivity to propose a dynamic presentation of web questionnaires. *Quality & Quantity*, *51*(3), 1321-1336.

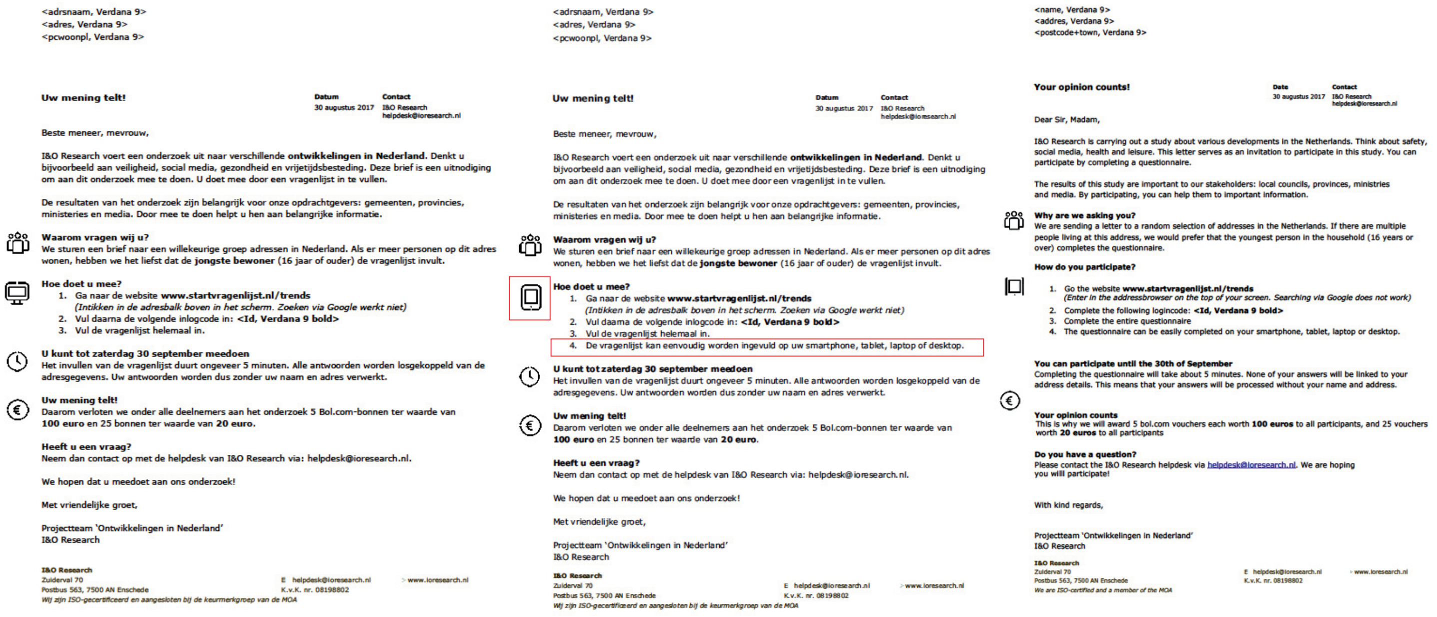
Roberts, A., de Leeuw, E.D., Hox, J., Klausch, T. L., & de Jongh, A. (2013). Leuker kunnen we het wel maken. Online vragenlijst design: standaard matrix of scrollmatrix? (We can make it nicer. Online questionnaire design: standard matrix or scrollmatrix). Jaarboek Marktonderzoeksassociatie, 133-149. Available on: http://moa04.artoo.nl/clou-moaweb-images/images/bestanden/pdf/Jaarboeken_MOA/JaarboekMarktonderzoek2013.pdf

Selkälä, A., & Couper, M. P. (2017). Automatic Versus Manual Forwarding in Web Surveys. *Social Science Computer Review*, doi: 10.1177/ 0894439317736831.

Statistics Netherlands (2017). StatLine: Internet: access, use and facilities. Available on https://opendata.cbs.nl/statline/#/CBS/nl/dataset/83429NED/table?dl=A399

Appendix A

Original Dutch versions of old and new invitation letter (and English translation of new letter)



<adrsnaam, Verdana 9>
<adres, Verdana 9>
<pcwoonpl, Verdana 9>

Uw mening telt!

Datum 30 augustus 2017
Contact I&O Research
helpdesk@ioresearch.nl

Beste meneer, mevrouw,

I&O Research voert een onderzoek uit naar verschillende **ontwikkelingen in Nederland**. Denkt u bijvoorbeeld aan veiligheid, social media, gezondheid en vrijetijdsbesteding. Deze brief is een uitnodiging om aan dit onderzoek mee te doen. U doet mee door een vragenlijst in te vullen.

De resultaten van het onderzoek zijn belangrijk voor onze opdrachtgevers: gemeenten, provincies, ministeries en media. Door mee te doen helpt u hen aan belangrijke informatie.

Waarom vragen wij u?
We sturen een brief naar een willekeurige groep adressen in Nederland. Als er meer personen op dit adres wonen, hebben we het liefst dat de **jongste bewoner** (16 jaar of ouder) de vragenlijst invult.

Hoe doet u mee?

1. Ga naar de website **www.startvragenlijst.nl/trends**
   (Intikken in de adresbalk boven in het scherm. Zoeken via Google werkt niet)
2. Vul daarna de volgende inlogcode in: **<Id, Verdana 9 bold>**
3. Vul de vragenlijst helemaal in.

U kunt tot zaterdag 30 september meedoen
Het invullen van de vragenlijst duurt ongeveer 5 minuten. Alle antwoorden worden losgekoppeld van de adresgegevens. Uw antwoorden worden dus zonder uw naam en adres verwerkt.

Uw mening telt!
Daarom verloten we onder alle deelnemers aan het onderzoek 5 Bol.com-bonnen ter waarde van **100 euro** en 25 bonnen ter waarde van **20 euro**.

Heeft u een vraag?
Neem dan contact op met de helpdesk van I&O Research via: helpdesk@ioresearch.nl.

We hopen dat u meedoet aan ons onderzoek!

Met vriendelijke groet,

Projectteam 'Ontwikkelingen in Nederland'
I&O Research

I&O Research
Zuiderval 70
Postbus 563, 7500 AN Enschede
Wij zijn ISO-gecertificeerd en aangesloten bij de keurmerkgroep van de MOA
E helpdesk@ioresearch.nl
K.v.K. nr. 08198802
> www.ioresearch.nl

<adrsnaam, Verdana 9>
<adres, Verdana 9>
<pcwoonpl, Verdana 9>

Uw mening telt!

Datum 30 augustus 2017
Contact I&O Research
helpdesk@ioresearch.nl

Beste meneer, mevrouw,

I&O Research voert een onderzoek uit naar verschillende **ontwikkelingen in Nederland**. Denkt u bijvoorbeeld aan veiligheid, social media, gezondheid en vrijetijdsbesteding. Deze brief is een uitnodiging om aan dit onderzoek mee te doen. U doet mee door een vragenlijst in te vullen.

De resultaten van het onderzoek zijn belangrijk voor onze opdrachtgevers: gemeenten, provincies, ministeries en media. Door mee te doen helpt u hen aan belangrijke informatie.

Waarom vragen wij u?
We sturen een brief naar een willekeurige groep adressen in Nederland. Als er meer personen op dit adres wonen, hebben we het liefst dat de **jongste bewoner** (16 jaar of ouder) de vragenlijst invult.

Hoe doet u mee?

1. Ga naar de website **www.startvragenlijst.nl/trends**
   (Intikken in de adresbalk boven in het scherm. Zoeken via Google werkt niet)
2. Vul daarna de volgende inlogcode in: **<Id, Verdana 9 bold>**
3. Vul de vragenlijst helemaal in.
4. De vragenlijst kan eenvoudig worden ingevuld op uw smartphone, tablet, laptop of desktop.

U kunt tot zaterdag 30 september meedoen
Het invullen van de vragenlijst duurt ongeveer 5 minuten. Alle antwoorden worden losgekoppeld van de adresgegevens. Uw antwoorden worden dus zonder uw naam en adres verwerkt.

Uw mening telt!
Daarom verloten we onder alle deelnemers aan het onderzoek 5 Bol.com-bonnen ter waarde van **100 euro** en 25 bonnen ter waarde van **20 euro**.

Heeft u een vraag?
Neem dan contact op met de helpdesk van I&O Research via: helpdesk@ioresearch.nl.

We hopen dat u meedoet aan ons onderzoek!

Met vriendelijke groet,

Projectteam 'Ontwikkelingen in Nederland'
I&O Research

I&O Research
Zuiderval 70
Postbus 563, 7500 AN Enschede
Wij zijn ISO-gecertificeerd en aangesloten bij de keurmerkgroep van de MOA
E helpdesk@ioresearch.nl
K.v.K. nr. 08198802
> www.ioresearch.nl

<name, Verdana 9>
<addres, Verdana 9>
<postcode+town, Verdana 9>

Your opinion counts!

Date 30 augustus 2017
Contact I&O Research
helpdesk@ioresearch.nl

Dear Sir, Madam,

I&O Research is carrying out a study about various developments in the Netherlands. Think about safety, social media, health and leisure. This letter serves as an invitation to participate in this study. You can participate by completing a questionnaire.

The results of this study are important to our stakeholders: local councils, provinces, ministries and media. By participating, you can help them to important information.

Why are we asking you?
We are sending a letter to a random selection of addresses in the Netherlands. If there are multiple people living at this address, we would prefer that the youngest person in the household (16 years or over) completes the questionnaire.

How do you participate?

1. Go the website **www.startvragenlijst.nl/trends**
   (Enter in the addressbrowser on the top of your screen. Searching via Google does not work)
2. Complete the following logincode: **<Id, Verdana 9 bold>**
3. Complete the entire questionnaire
4. The questionnaire can be easily completed on your smartphone, tablet, laptop or desktop.

You can participate until the 30th of September
Completing the questionnaire will take about 5 minutes. None of your answers will be linked to your address details. This means that your answers will be processed without your name and address.

Your opinion counts
This is why we will award 5 bol.com vouchers each worth **100 euros** to all participants, and 25 vouchers worth **20 euros** to all participants

Do you have a question?
Please contact the I&O Research helpdesk via helpdesk@ioresearch.nl. We are hoping you willl participate!

With kind regards,

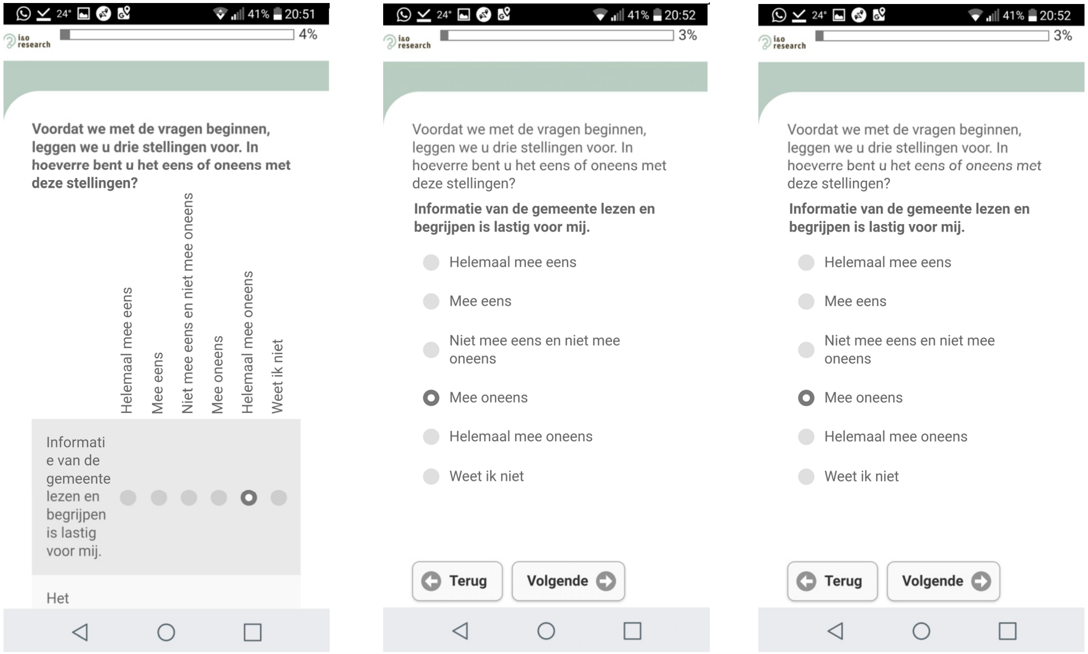
Projectteam 'Ontwikkelingen in Nederland'
I&O Research

I&O Research
Zuiderval 70
Postbus 563, 7500 AN Enschede
We are ISO-certified and a member of the MOA
E helpdesk@ioresearch.nl
K.v.K. nr. 08198802
> www.ioresearch.nl

Figure A1 Old invitation letter (left) in conditions 1 and 3, new letter (middle) used in conditions 2, 4 and 5, and English translation (right) of new letter. The difference between the letters is emphasized by the red boxes, which were not included in the actual letter

Appendix B

Screenshots of question 1 from the survey across the different layout conditions



Voordat we met de vragen beginnen, leggen we u drie stellingen voor. In hoeverre bent u het eens of oneens met deze stellingen?

	Helemaal mee eens	Mee eens	Niet mee eens en niet mee oneens	Mee oneens	Helemaal mee oneens	Weet ik niet
Informatie van de gemeente lezen en begrijpen is lastig voor mij.	○	○	○	○	●	○
Het						

Voordat we met de vragen beginnen, leggen we u drie stellingen voor. In hoeverre bent u het eens of oneens met deze stellingen?

Informatie van de gemeente lezen en begrijpen is lastig voor mij.

○ Helemaal mee eens
○ Mee eens
○ Niet mee eens en niet mee oneens
● Mee oneens
○ Helemaal mee oneens
○ Weet ik niet

Voordat we met de vragen beginnen, leggen we u drie stellingen voor. In hoeverre bent u het eens of oneens met deze stellingen?

Informatie van de gemeente lezen en begrijpen is lastig voor mij.

○ Helemaal mee eens
○ Mee eens
○ Niet mee eens en niet mee oneens
● Mee oneens
○ Helemaal mee oneens
○ Weet ik niet

Figure B1 Screenshots of question 1 for survey in old design (conditions 1 - left), new design (condition 2 and 3 - middle) and new design + autoforward (conditions 4 and 5 - right) taken on a LG G6. The only difference between the middle and right panel is in how the respondent navigates (manually vs. autoforwarding.

methods, data, analyses | Vol. 13(2), 2019, pp. 307-320 DOI: 10.12758/mda.2019.03

Sensitive Question Techniques and Careless Responding: Adjusting the Crosswise Model for Random Answers

Patrick Schnapp

Abstract

The crosswise model is a popular sensitive question technique often considered more accurate than direct questioning. When this technique is used, a sensitive question is paired with a nonsensitive question that has a known prevalence and respondents are asked to give a joint answer to the pair of questions. Recent research has shown that prevalence estimates based on the crosswise model are biased towards 50% when respondents answer randomly, and that random answers are frequent. I develop methods to adjust the crosswise model for self-reported random answers. Results from an exploratory online survey (n = 103) show that (i) fewer respondents report random answers than might be expected given unadjusted results, (ii) results differ considerably between questions, and (iii) one of three questions yields an estimate that is substantially and significantly above the true value even after adjusting for random answers.

Keywords: Crosswise model; randomized response technique; sensitive question techniques; socially desirable responding; careless responding; random answers; satisficing

Survey researchers are often interested in answers to sensitive questions, i.e., "questions about violations of social norms by the respondent's behavior" (Näher & Krumpal, 2012, p. 1603).[1] When answering such questions, some respondents exhibit socially desirable responding, "the tendency to give overly positive self-descriptions" (Paulhus, 2002, p. 50). For example, many respondents who have engaged in drunk driving will not admit this in a survey (Locander, Sudman, & Bradburn, 1976). This entails two problems. First, researchers who take respondents' answers at face value will underestimate the prevalence of undesirable characteristics and overestimate the prevalence of desirable characteristics. Second, researchers' ability to estimate associations between the characteristic in question and other variables will be impeded, as the respondents' answers are influenced by both their true status on the characteristic and their tendency to engage in socially desirable responding (Wolter & Preisendörfer, 2013).

One approach to this problem is the use of sensitive question techniques. These allow the respondent to hide his or her answer to the sensitive question, but also allow the researcher to estimate the prevalence of the sensitive characteristic in the sample as a whole, as well as associations between the sensitive and other characteristics.

A technique recently popular is the crosswise model (CM), a variant of the randomized response technique (Warner, 1965) introduced by Yu, Tian, and Tang (2008). In the CM, respondents are asked to reply to a combination of two yes/no questions. One is the sensitive question, the other is nonsensitive. For example, the respondent may be asked (i) whether he or she has ever engaged in drunk driving and (ii) whether his or her mother was born in January, February or March. The respondent is then asked whether (A) the answer to the two questions is the same (both "yes" or both "no") or (B) the answers to the questions are different. The prevalence of the sensitive characteristic (drunk driving) can be estimated because the prevalence of the nonsensitive item (mother's month of birth) is known (approx. .25).

Prevalence estimates on the basis of the CM are often significantly higher than those derived from direct questions (Enzmann, 2017; Hoffmann et al., 2015; Hoffmann & Musch, 2016, 2018; Höglinger & Jann, 2018; Höglinger, Jann, &

[1] The term "sensitive question" is also used for related but different concepts (for in-depth discussions, Krumpal, 2013; Tourangeau & Yan, 2007).

Acknowledgments

I thank Thomas Ullmann of maQ for technical support, all survey respondents for participating, and two reviewers for helpful comments.

Direct correspondence to

Patrick Schnapp
E-mail: p_schnapp@gmx.de

Diekmann, 2016; Hopp & Speil, 2019; Jann, Jerke, & Krumpal, 2012; Korndörfer, Krumpal, & Schmukle, 2014; Kundt, 2014; Waubert de Puiseau, Hoffmann, & Musch, 2017). Many researchers interpret this as evidence for the superiority of the CM (e.g., Hopp & Speil, 2019; Kazemzadeh et al., 2016; Kundt et al., 2017; Waubert de Puiseau et al., 2017). These authors rely on the more-is-better assumption, according to which techniques that yield higher prevalence estimates of sensitive characteristics are more valid. The assumption is unwarranted in the case of the CM. This is because the more respondents choose an answer at random, the more the prevalence estimate will be biased towards 50% (Enzmann, 2017). Random answers will hence bias estimates downwards when the true prevalence is above 50% and upwards when the true prevalence is below 50%, as is often the case with sensitive characteristics (Höglinger & Diekmann, 2017). While the same is true of direct questions (Hemenway, 1997), the fact that CM questions are more complex makes it more likely that respondents will answer randomly because they are unwilling or unable to put in the cognitive effort necessary for choosing the correct response. Thus, the more-is-better assumption may lead to the conclusion that the CM produces more valid results than direct questions, when in fact the higher estimates are a consequence of random responding (Höglinger & Diekmann, 2017).

Three recent studies shine a light on this issue. Höglinger and Diekmann (2017) asked respondents whether they had ever received a donated organ and whether they had ever suffered from Chagas disease. The prevalence of both characteristics was assumed to be zero. Under the assumption of no other causes for bias, the rate of random answers is twice the false positive rate. After removing cases based on apparently problematic nonsensitive items, prevalence estimates were 6% (organ) and 1% (Chagas), implying random answer rates of 12% and 2%, respectively. Höglinger and Jann (2018), studying cheating in games, estimated that the CM misclassified 14% of respondents, implying 28% answered randomly under the same assumption. Enzmann (2017) reported results from a survey asking students to report illegal behavior. In a follow-up to one of the CM questions, respondents were asked how they had answered the question, with 13% stating they had answered randomly.

There also are studies showing that standard CM prevalence estimates are close to the true values (Hoffmann et al., 2015; Hoffman & Musch, 2016, 2018). It is important to appreciate what this does and does not show. These studies are evidence in favor of the accuracy of the CM as a measure of prevalence. It is unknown, however, whether individual respondents in these studies answered truthfully or not. In aggregate estimates, incorrect answers in both directions may even each other out to produce an estimate close to the true value (Höglinger & Diekmann, 2017). Incorrect answers at the individual level impede researchers' ability to correctly estimate the association between the sensitive characteristic and other variables (Enzmann, 2017), even if aggregate prevalence estimates are correct. The appeal of Höglinger & Diekmann's (2017) validation strategy of using zero-prevalence items is that it allows the researcher to estimate the rate of false positives even if the distribution of the sensitive characteristic cannot be measured.

This body of research suggests a potential remedy to the problem of bias due to random answers, and a way of testing its validity. In a survey, CM questions may be followed by direct questions asking whether the respondent answered the CM question randomly. Adjusted prevalence estimates can be calculated. These estimates can be compared to the unadjusted estimates and known true values. The present paper reports the results of a small, exploratory study to demonstrate the application of the technique and present first results.

The remainder of this article is organized as follows. In the theoretical part of the paper, I review formulae for the standard crosswise model and variants that feature adjustments. In the empirical part, I report results of the exploratory survey, showing estimates on the basis of different versions of the CM. Finally, results are discussed.

Crosswise Estimation

The Standard Crosswise Model

Suppose we are interested in the prevalence of a sensitive behavior, such as drunk driving. We could present respondents with the question about the respondents' drunk driving and couple it with a question about a nonsensitive matter, such as whether the respondent's mother's birthday is between January and March. We then ask whether (A) the answer to the two questions is the same (both "yes" or both "no") or (B) the answers to the two questions is different. When the standard CM is applied, the estimator of the prevalence of the sensitive characteristic is (Yu et al., 2008, notation altered)

$$\hat{\pi}_{SCM} = \frac{\hat{\lambda} + p - 1}{2p - 1}; p \neq 0.5 \tag{1}$$

where $\hat{\pi}_{SCM}$ is the estimate of the proportion of respondents carrying the sensitive characteristic estimated by the standard CM, $\hat{\lambda}$ is the estimate of the proportion of respondents whose true answer is "A" ("the same"), and p is the proportion of respondents for whom the true answer to the nonsensitive question is "yes". The proportion of respondents for whom the true answer is "A" is estimated as (Yu et al., 2008, notation altered)

$$\hat{\lambda} = \frac{n_A}{n} \tag{2}$$

where n_A is the sum of "A" answers and n is the sample size.

An unbiased estimator of the variance of $\hat{\pi}_{SCM}$ is (Yu et al., 2008)

$$\bar{\mathbb{V}}\left[\hat{\pi}_{SCM}\right]=\frac{\hat{\lambda}\left(1-\hat{\lambda}\right)}{(n-1)(2p-1)^2}=\frac{\hat{\pi}_{SCM}\left(1-\hat{\pi}_{SCM}\right)}{(n-1)}+\frac{p(1-p)}{(n-1)(2p-1)^2}\,;\, p\neq 0.5 \qquad (3)$$

The right-hand side of the equation shows the decomposition of the variance into the sampling part and an additional term due to the uncertainty introduced by the use of the nonsensitive item (Kundt, 2014).

Adjusting the Crosswise Model for Random Answers

Some respondents may answer CM questions randomly (choosing either answer with equal probability). As can be seen from formulae (2) and (1), this biases $\hat{\lambda}$. and hence $\hat{\pi}_{SCM}$ toward 0.5. However, we may be able to estimate the proportion of CM questions that were answered randomly (e.g., on the basis of follow-up questions). Then estimates adjusted for random answers may be derived by (Enzmann, 2017, notation altered)

$$\hat{\pi}_{CMR-S}=\frac{\hat{\pi}_{SCM}-0.5r}{1-r}\,;\, r\neq 1 \qquad (4)$$

where *CMR-S* stands for "crosswise model adjusted for random answers at the level of the sample" and r is the proportion of random answers. When $r = 1$, the result is undefined. This is as it should be; if all respondents answer randomly, $\hat{\pi}_{SCM}$ carries no information about the true value of π.

Equation (4) implies at the variance may be calculated as

$$\bar{\mathbb{V}}\hat{\pi}_{CMR-S}=\frac{\bar{\mathbb{V}}\left[\hat{\pi}_{SCM}\right]}{(1-r)^2}=\left[\frac{\hat{\pi}_{SCM}\left(1-\hat{\pi}_{SCM}\right)}{(n-1)}+\frac{p(1-p)}{(n-1)(2p-1)^2}\right]/\left(1-r\right)^2;r\neq\ 1 \qquad (5)$$

This variance is hence larger than the variance of the standard crosswise model if $r > 0$, reflecting the added uncertainty introduced by random answers.

However, if information about random answering can be linked to individual respondents' answers to individual items, this can be taken into account more directly in what I call the CMR-I (crosswise model adjusted for random answers at the individual level). If the respondent stated that he answered randomly, then the value for A should be set to 0.5; if he did not, the value should remain unchanged (i.e., 1 for an "A" answer and 0 for a "B" answer). Formally,

$$A_{adj} = 0.5R + (Y = A)(1 - R) \qquad (6)$$

where A_{adj} is the adjusted value for the A variable at the individual level, R is an indicator variable taking the value 1 if the respondent answered the question randomly and $Y = A$ is the unadjusted value for the answer to the CM question, taking the value 1 for an "A" answer and 0 for a "B" answer. As can be seen from the formula, both "A" answers ($A = 1$) and "B" answers ($A = 0$) are converted to 0.5 if the respondent answered the question randomly ($R = 1$); otherwise ($R = 0$), they remain unchanged. $n_{A_{adj}}$ (the sum of the A_{adj} variable) can then be used instead of n_A (the sum of the unadjusted variable) in (2). In the crosswise model adjusted for random answers at the individual level, the estimate is hence given by (1), (6) and

$$\hat{\lambda} = \frac{n_{A_{adj}}}{n} \tag{7}$$

The variance is given by (3) rather than (5) under the assumption that information on random answering is correct.

Data and Methods

An online survey was conducted. The German-language questionnaire was designed to produce a sufficient number of random answers to test whether adjusting for them leads to improvements in the estimate, but no attempt was made to actively confuse participants. After an introductory page, participants received instructions on how to answer CM questions (but no practice examples). This was followed by the main part of the questionnaire. Following Höglinger and Diekmann (2017), lifetime prevalences of three rare diseases were chosen as zero-prevalence items (Castleman disease, Chagas disease, Barth syndrome). Similar to Diekmann (2012) and Kundt (2014), one nonsensitive item asked about the respondent's house number; the other two are standard nonsensitive questions in the CM literature (concerning mother's and father's month of birth being January or February). Each crosswise question was followed by a companion question on the next page. Respondents were informed that many participants find these types of questions hard to answer and asked whether they had "just chosen an answer at random" (answers: yes/no). English translations of all CM and companion questions are displayed in the appendix. Sociodemographic information was also collected. The last question asked whether respondents had answered this survey before; this was accompanied by the information that their answer would have no influence on their obtaining the code they needed to gain points (see below). CM and companion questions were obligatory, other questions were optional. Respondents had to click through to the last page of the questionnaire to obtain the code.

The survey was programmed in maQ (Ullmann, 2004) and posted on two sites, Survey Circle and Poll Pool. On both, members can fill in surveys to gain points. The more points a member has, the more points other members can gain when filling in his or her survey. The questionnaire's last page contained codes necessary to obtain points. The questionnaire was advertised as a "Short profile on health", open to all participants who were at least 18 years old and had passed their last school exam in Germany. Answers were collected in October and November 2018.

Data on months of birth in Germany were obtained directly from the Federal Statistical Office and date back to 1948. The distribution of first digits of house numbers was taken from Kundt (2014). Age was approximated by subtracting the year of birth from 2018. Answers to CM questions were set to missing if their companion questions had not been answered.

Results

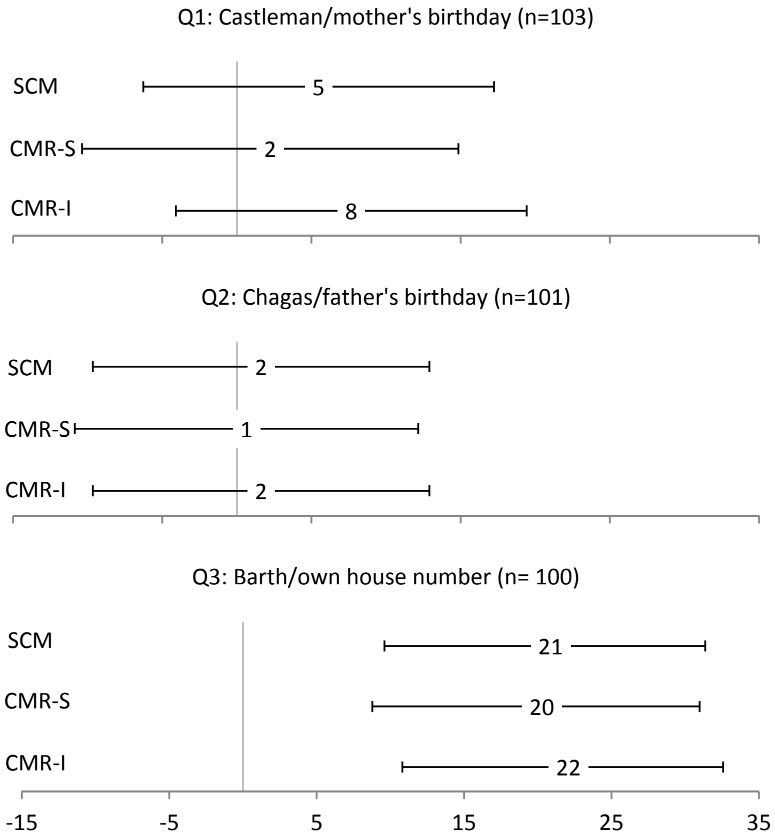
One hundred and nine respondents answered at least one combination of CM and companion question. Six respondents were excluded because they stated they had answered the survey before or were not sure. The sample size is hence 103, but there is some missing data. Descriptive statistics are shown in Table 1. The youngest respondents were born in 1998. Under the assumption that parents would have been at least 20 years old when the respondents were born, the proportion of parents born in January or February was calculated for the years 1948 to 1978; the result is approx. 16.7 percent. The proportion of German house numbers starting with the digit 8 or 9 is approx. 8.8% (Kundt, 2014).

Table 1 Descriptive statistics.

	Mean / Proportion	*SD*	Minimum	Maximum	*n*
Gender (1=male)	0.34	0.05	0	1	98
Passed "Abitur" exam	0.91	0.03	0	1	100
Year of Birth	1991.64	5.79	1962	1998	99
Age	26.36	5.79	18	54	99
Answered randomly					
Q1	0.07	0.02	0	1	103
Q2	0.02	0.01	0	1	101
Q3	0.02	0.01	0	1	100

Abbreviations. SD: standard deviation; Q: question

Figure 1 Estimates of the percentages of respondents carrying the sensitive characteristic



Abbreviations. SCM: standard crosswise model; CMR-S: crosswise model adjusted for random answers at the sample level; CMR-I: crosswise model adjusted for random answers at the individual level.

The proportions of self-reported random answers are low and differ considerably between questions. They are 6.8% for Q3 (question 3), 2.0% for Q2, and 2.0% for Q3.

Figure 1 presents point estimates and 95% confidence intervals for all combinations of questions and types of CM. In all cases, an unbiased point estimate would be zero. All point estimates are above zero, but the size of the bias differs considerably between questions. Most strikingly, answers to Q3 depart substantially and significantly from the true value. Under the assumption of no other causes for bias, this implies that 42 percent of respondents answered Q3 randomly.

These question effects are larger overall than the effects of the estimation method. The CMR-I improves on the standard estimates in no case and does worse in two. In contrast, the CMR-S fares a little better than both the standard method and the CMR-I in all cases. These differences are far from significant, however.

Discussion

Results from this small study show that point estimates based on the standard CM are higher than the true value; in one case, the difference is very large and statistically significant. This result adds to validation studies showing that the more-is-better assumption is invalid in the case of the CM. It also casts doubt on results from the literature in which the CM was applied. If readers consult such studies, they could apply a mental correction for random responding using the formulae given above and reasonable assumptions about the proportion of responses that are random. In this context, note that the present results were obtained from respondents who were extrinsically motivated to participate. A substantial proportion of respondents probably participated despite low intrinsic interest, and Brower (2018) reports negative associations between respondent interest and measures of careless responding. It hence seems likely that the bias observed in this study, as well as other CM studies using incentives for participation, is higher than it would have been if respondents had been intrinsically motivated to participate.

One may wonder why the results are so different for Q3. Possible reasons include (i) the position of the question, (ii) the content of the sensitive item (perhaps respondents mistook Barth syndrome for something else), (iii) the prevalence of the nonsensitive item; (iv) the person the nonsensitive question referred to (self), and (v) the topic of the nonsensitive question (house number).

The position of the sensitive question may seem unlikely to have played a large role given that Höglinger and Diekmann (2017) found no substantial or significant positional effects. However, it is possible that by the time they reached Q3, some respondents were sufficiently disappointed by the contents of the survey to start giving random answers. Respondents who start a survey advertised as a "Short profile on health" may expect questions concerning their exercise and eating habits rather that questions about rare diseases in an unusual question format. Such an effect could be particularly strong if respondents who live a healthy lifestyle self-selected into the survey because they were looking forward to presenting themselves in a favorable light, an opportunity not given by the questionnaire. Some of these respondents might have combined random answers to Q3 with an untruthful answer to the companion question.

Concerning the content of the sensitive item, it is unclear how the respondents might have misunderstood what “Barth syndrome” refers to.

While the prevalence of the nonsensitive item in Q3 is low and hence accords the respondent little protection of his privacy, Diekmann (2012) showed that the students in his sample overestimated the proportion of house numbers starting in high digits. However, it is conceivable that respondents were unwilling to answer truthfully to a question involving their own house number in a time in which data security had been a prominent topic in the German media due to the introduction of the new General Data Protection Regulation.

There is no definitive answer to the question why Q3 performed so much worse than the other two. To avoid such uncertainty, future researchers wishing to test the CMR-I may prefer to vary features of sensitive and nonsensitive questions randomly.

The main result is that adjusting for random answering, as implemented in this study, does little to remove bias from CM results. A number of potential explanations present themselves. First, the companion question did not ask about deliberately false answers. Hence, the study was not designed to remove bias from this source, if any. Second, the companion questions themselves were sensitive, as answering randomly in a survey violates the norm of honesty. This may lead to socially desirable responding to these questions, impeding sufficient adjustments. Third, a yes/no question is too crude to measure random answering. A scale with more than two points may be preferable, as such scales generally exhibit better psychometric properties than binary scales (Krosnick & Presser, 2010; Markon, Chmielewski, & Miller, 2011) and can measure degrees of certainty that the correct answer was given. Ideally, such a question would also capture the fact that some respondents intentionally try to give the wrong answer, but may not be sure whether they succeeded in doing so. Authors who would like to pursue this avenue of research may also want to test whether it is really helpful to ask a companion question after every CM question – a design feature that seems impractical for applied surveys. Results may be equally good or better if the questionnaire presents a battery of CM questions followed by a summary companion question asking respondents what proportion of CM questions they answered randomly.

These may be fruitful avenues for future research. While the results of this and other studies suggest that the crosswise model has shortcomings, the problem of socially desirable responding is too serious to give up on techniques that may lead to viable solutions after all.

References

Brower, C. K. (2018). *Too long and too boring: The effects of survey length and interest on careless responding* (Master's Thesis). Fairborn: Wayne State University.

Diekmann, A. (2012). Making use of ''Benford's law'' for the randomized response technique. *Sociological Methods and Research*, *41*(2), 325-334. https://doi.org/10.1177/0049124112452525

Enzmann, D. (2017). Die Anwendbarkeit des Crosswise-Modells zur Prüfung kultureller Unterschiede sozial erwünschten Antwortverhaltens: Implikationen für seinen Einsatz in internationalen Studien zu selbstberichteter Delinquenz. In S. Eifler & F. Faulbaum (Eds.), *Methodische Probleme von Mixed-Mode-Ansätzen in der Umfrageforschung* (pp. 231-269). Wiesbaden: VS.

Hemenway, D. (1997). The myth of millions of annual self-defense gun uses: A case study of survey overestimates of rare events. *Chance, 10*(3), 6-10. https://doi.org/10.1080/09332480.1997.10542033

Hoffmann, A., Diedenhofen, B., Verschuere, B., & Musch, J. (2015). A strong validation of the crosswise model using experimentally-induced cheating behavior. *Experimental Psychology, 40*(6), 403-414. https://doi.org/10.1027/1618-3169/a000304

Hoffmann, A., & Musch, J. (2016). Assessing the validity of two indirect questioning techniques: A stochastic lie detector versus the crosswise model. *Behavior Research Methods, 48*(3), 1032-1046. https://doi.org/10.3758/s13428-015-0628-6

Hoffmann, A., & Musch, J. (2018). Prejudice against women leaders: Insights from an indirect questioning approach. *Sex Roles*. Advance online publication. https://doi.org/10.1007.s11199-018-0969-6

Höglinger, M., & Diekmann, A. (2017). Uncovering a blind spot in sensitive question research: False positives undermine the crosswise-model RRT. *Political Analysis, 25*(1), 131-137. https://doi.org/10.1017/pan.2016.5

Höglinger, M., & Jann, B. (2018). More is not always better: An experimental individual-level validation of the randomized response technique and the crosswise model. *PLoS ONE, 13*(8), e0201770. https://doi.org/10.1371/journal.pone.0201770

Höglinger, M., Jann, B., & Diekmann, A. (2016). Sensitive questions in online surveys: An experimental evaluation of different implementations of the randomized response technique and the crosswise model. *Survey Research Methods, 10*(3), 171-187. https://doi.org/10.18148/srm/2016.v10i3.6703

Hopp, C., & Speil, A. (2019). Estimating the extent of deceitful behaviour using crosswise elicitation models. *Applied Economics Letters, 26*(5), 396-400. https://doi.org/10.1080/13504851.2018.1486007

Jann, B., Jerke, J., & Krumpal, I. (2012). Asking sensitive questions using the crosswise model: An experimental survey measuring plagiarism. *Public Opinion Quarterly, 76*(1), 32-49. https://doi.org/10.1093/poq/nfr036

Kazemzadeh, Y., Shokoohi, M., Baneshi, M. R., & Haghdoost, A. A. (2016). The frequency of high-risk behaviors among Iranian college students using indirect methods: Network scale-up and crosswise model. *International Journal of High Risk Behaviors & Addiction, 5*(3), e25130. https://doi.org/10.5812/ijhrba.25130

Korndörfer, M., Krumpal, I., & Schmukle, S. C. (2014). Measuring and explaining tax evasion: Improving self-reports using the crosswise model. *Journal of Economic Psychology, 45*(1), 18-32. https://doi.org/10.1016/j.joep.2014.08.001

Krosnick, J. A., & Presser, S. (2010). Question and questionnaire design. In P. V. Marsden & J. D. Wright (eds.), *Handbook of Survey Research* (pp. 263-313). Bingley: Emerald.

Krumpal, I. (2013). Determinants of social desirability bias in sensitive surveys: A literature review. *Quality and Quantity, 47*(4), 2025–2047. https://doi.org/10.1007/s11135-011-9640-9

Kundt, T. (2014). *Applying "Benford's law" to the crosswise model: Findings from an online survey on tax evasion* (Working Paper Series No. 148). Hamburg: Helmut Schmidt Universität.

Kundt, T. C., Misch, F., & Nerré, B. (2017). Re-assessing the merits of measuring tax evasion through business surveys: An application of the crosswise model. *International Tax and Public Finance, 24*(1), 112-133. https://doi.org/10.1007/s10797-015-9373-0

Locander, W., Sudman, S., & Bradburn, N. (1976). An investigation of interview method, threat and response distortion. *Journal of the American Statistical Association, 71*(354), 269-275. https://doi.org/10.1080/01621459.1976.10480332

Markon, K. E., Chmielewski, M., & Miller, C. J. (2011). The reliability and validity of discrete and continuous measures of psychopathology: A quantitative review. *Psychological Bulletin, 137*(4), 856-879. https://doi.org/10.1037/a0023678

Näher, A.-F., & Krumpal, I. (2012). Asking sensitive questions: The impact of forgiving wording and question context on social desirability bias. *Quality & Quantity, 46*(5), 1601-1616. https://doi.org/10.1007/s11135-011-9469-2

Tourangeau, R., & Yan, T. (2007). Sensitive questions in surveys. *Psychological Bulletin, 133*(5), 859-883. https://doi.org/10.1037/0033-2909.133.5.859

Ullmann, T. D. (2004). *maQ-Fragebogengenerator: Make a questionnaire*. Retrieved from http://maq-online.de

Warner, S. L. (1965). Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association, 60*(309), 63-69. https://doi.org/10.1080/01621459.1965.10480775

Waubert de Puiseau, B., Hoffmann, A., & Musch, J. (2017). How indirect questioning techniques may promote democracy: A preelection polling experiment. *Basic and Applied Social Psychology, 39*(4), 209-217. https://doi.org/10.1080/01973533.2017.1331351

Wolter, F., & Preisendörfer, P. (2013). Asking sensitive questions: An evaluation of the randomized response technique versus direct questioning using individual validation data. *Sociological Methods & Research, 42*(3), 321-353. https://doi.org/10.1177/0049124113500474

Yu, J. W., Tian, G. L., & Tang, M. L. (2008). Two new models for survey sampling with sensitive characteristic: Design and analysis. *Metrika, 67*(3), 251-263. https://doi.org/10.1007/s00184-007-0131-x

Appendix A

Question wordings and answer options for the crosswise model

Page 3

Here are two questions:

A: Was your **mother** born in January or February?

B: Have you ever been diagnosed with Castleman disease?

Page 5

Here are two questions:

A: Was your **father** born in January or February?

B: Have you ever been diagnosed with Chagas disease (a.k.a. American trypanosomiasis)?

Page 7

Here are two questions:

A: Please think of your **main residence**. Is the **first digit** of your house number 8 or 9?

B: Have you ever been diagnosed with Barth syndrome?

Item on pages 3, 5, 7

Which of the following statements is true?

(Answers: The answer to both questions is the same (twice yes or twice no) / The answers to the two questions are different (once yes and once no, irrespective of the order))

Pages 4. 6, 8 (companion question)

Many respondents find he type of question you have just answered hard.

Is the following statement correct?

Answering the question on the previous page, I just chose an answer at random. (Answers: yes / no)

methods, data, analyses | Vol. 13(2), 2019, pp. 321-340 DOI: 10.12758/mda.2019.02

Methodological Aspects of a Quantitative and Qualitative Survey of Asylum Seekers in Germany – A Field Report

Sonja Haug[1], *Susanne Lochner*[2] *& Dominik Huber*[1]

[1] *Ostbayerische Technische Hochschule Regensburg*
[2] *German Youth Institute (DJI) Munich*

Abstract

This field report presents and discusses methodological issues and challenges encountered in a mixed-methods research project on asylum seekers in Bavaria, Germany. It documents the research design of, and field experiences in, a quantitative survey based on a quota sampling procedure and a qualitative study, both of which were conducted in collective accommodation for asylum seekers at selected locations in that federal state. Standardized PAPI multiple-topic questionnaires were completed by asylum seekers from Syria, Afghanistan, Eritrea, and Iraq ($N = 779$); most of the questionnaires were self-administered. In addition, 12 qualitative face-to-face biographical interviews were conducted in order to gain an in-depth understanding of attitudes and experiences of asylum seekers. This report focuses on the following aspects: the use of gatekeepers to facilitate participant recruitment; sampling procedures; the involvement of interpreters in the data collection process; response bias and response behaviors among asylum seekers; and the experiences gained from data collection in collective accommodation for asylum seekers.

Keywords: asylum seekers in Germany, mixed methods, surveys, biographical research, gatekeeper approach; methodological issues; use of interpreters

In 2015 and 2016, about 1.16 million asylum seekers were registered in Germany (BAMF, 2018, p. 3). During this period – and due to the distribution of asylum seekers among the German federal states according to a quota system known as the “Königstein key” (*Königsteiner Schlüssel*) – Bavaria was allocated 15.33% of persons seeking asylum in Germany. In absolute figures, this meant that 67,639 first-time asylum applications were registered in Bavaria in 2015 (BAMF, 2016b, p. 16) and 82,003 in 2016 (BAMF, 2017a, p. 16), resulting in a total of almost 150,000 registrations. The increased number of asylum seekers in Germany, and the societal and political implications thereof, heightened the need for empirical data on these new arrivals. In 2016, the Ostbayerische Technische Hochschule Regensburg (OTH Regensburg) initiated a mixed-methods pilot study entitled “Asylsuchende in Bayern” (Asylum Seekers in Bavaria; see Haug, Currle, Lochner, Huber, & Altenbuchner, 2017) on behalf of the Hanns Seidel Foundation in order to gain a better understanding of the asylum seekers who arrived in that federal state in 2015 and 2016. The objective of this study was to enhance understanding of the motivations, sociodemographic characteristics, and attitudes of asylum seekers.

Prior to the qualitative and quantitative surveys carried out within the framework of the study, expert interviews were conducted with persons entrusted with the accommodation, distribution, and integration of asylum seekers. The findings of these interviews facilitated the design of a standardized quantitative questionnaire, which included questions on sociodemographic characteristics, value orientations, religiosity, and intentions to remain in Germany. Data collection was supported by interpreters, who delivered the questionnaires to asylum seekers in collective accommodation centers and were available to clarify in the respondents’ mother tongue any issues regarding the survey questions. In total, 779 asylum seekers participated in the quantitative survey.

In addition, 12 qualitative, face-to-face interviews were conducted with asylum seekers in order to collect exemplary biographies with the aim of gaining an in-depth understanding of their reasons for flight, value orientations, attitudes, and

Acknowledgments

We thank the audience of a presentation delivered at the Congress of the German Sociological Association in Bamberg in September 2016 and the two anonymous reviewers for their very helpful comments and suggestions. We also thank the research team (Amelie Altenbuchner, Edda Currle, the student assistants, and the native-speaker interpreters). The work was supported by the Hanns Seidel Foundation.

Direct correspondence to

Sonja Haug, Ostbayerische Technische Hochschule (OTH) Regensburg,
Fakultät für Angewandte Sozial- und Gesundheitswissenschaften, Seybothstraße 2,
93053 Regensburg
E-mail: sonja.haug@oth-regensburg.de

future aspirations. Time-line analysis and mapping was used to visualize the asylum seekers' biographies and the routes they took when they fled.

This article discusses methodological issues and challenges of collecting and analyzing quantitative and qualitative data from a vulnerable group, in this case asylum seekers. The following sections describe the procedures and research instruments used and discuss issues and challenges such as sampling, response bias, the use of interpreters and gatekeepers, data visualization, and research ethics.

Quantitative Methods

Sampling strategies for surveying asylum seekers in Germany: Issues and challenges

Standard methods for surveying persons with a migration background in Germany based on address and telephone registers were not applicable in this study. These methods include, first, the drawing of representative samples from addresses listed in municipal population registers. Register sampling requires cooperation with municipal statistics offices, which have access to the population registers. These registers contain addresses that could be used for postal or face-to-face surveys.

As in the case of migration background, the existence of a refugee background can be determined only in screening procedures after the survey. In the case of asylum seekers, the problem of collecting data on rare populations within the framework of a general population survey is particularly pronounced (Schnell et al., 2013a, pp. 285–288): An immensely large sample size would be needed in order to ensure a sufficient number of asylum seekers in the sample. Hence, this method was not a viable option in the present study.

A second method of surveying persons with a migration background in Germany entails drawing a disproportionate stratified municipal population register sample with prior classification of the population according to migration background using MigraPro (VDSt, 2013, p. 18; Haug et al., 2014, pp. 308–309). MigraPro is a German software tool that enables the classification of the population of persons with a migration background – especially first-generation migrants – in the municipal population registers based on citizenship and place of birth. However, as asylum status is not recorded in the municipal population registers, drawing a sample of asylum seekers using MigraPro would require an additional screening procedure within the random sample of persons with a migration background in order to identify asylum seekers. Hence, this method, too, was not feasible in the present study.

A third approach to surveying persons with a migration background in Germany is to use onomastic (i.e., name-based) methods to draw stratified samples.

The challenge here is to filter out on the basis of names as many persons as possible from certain countries of origin (Humpert & Schneiderheinze, 2002). As the onomastic method has the advantage of reduced screening effort, it is primarily used in Germany to pre-select migrants from certain countries of origin, for example Turkey or Poland, or from groups of countries, such as the former Soviet Union or former Yugoslavia, that have a shared history and a set of common typical names. The onomastic approach was used, for example, in Haug, Müssig, and Stichs' (2009) study on Muslims with a migration background in Germany, which considered almost 50 countries of origin with a predominantly Muslim population. The onomastic approach is more accurate for some countries of origin, such as Turkey, than for others, for example Russia (Schnell et al., 2013b). A test of the onomastic method using data from the German Socio-Economic Panel (SOEP) revealed that it could assign almost all available full names to a country-of-origin-of-the-language group (Liebau, Humpert, & Schneiderheinze, 2018).

The advantage of the onomastic method is that the language likely spoken by the respondents can be predicted on the basis of their names, so that interviewers with a knowledge of these heritage languages can be deployed for the process of data collection (Haug & Vernim, 2015). However, in the case of this method, too, prior screening would be necessary to determine the actual country of origin and the asylum status. Because we did not have access to any registers containing recent asylum seekers, it was not possible to employ the onomastic method in the present study.

The fourth method of surveying persons with a migration background in Germany is to draw samples of asylum seekers from the Central Register of Foreigners (Ausländerzentralregister, AZR), which is centrally managed by the Federal Office for Migration and Refugees (BAMF) in Nuremberg, Bavaria. The advantage of AZR-based sampling is that asylum status is available as a sampling characteristic – in addition to citizenship, sex, date of birth, year of entry, federal state (but not place) of residence, and the competent foreigners authority. However, as the AZR does not contain addresses, the competent foreigners authorities must be contacted individually in a further step in order to obtain the addresses of the persons in the sample (see Worbs, Bund, & Böhm, 2016 for a first application of this method). The IAB-BAMF-SOEP Survey of Refugees (Brücker, Rother, & Schupp, 2016c) and the qualitative preliminary study (Brücker et al., 2016a), which was conducted by a team including the BAMF research group, used the AZR pursuant to a statutory provision that came into effect on February 2, 2016 (§ 24a (5) of the Act on the Central Register of Foreigners, AZR).[1] Access to the AZR for university research purposes is subject to strict statutory restrictions and requires the permission of

[1] The transfer of data on foreigners from BAMF to research institutes is permitted if the data are needed to conduct collaborative scientific research pursuant to §75 (4) of the German Residence Act (Aufenthaltsgesetz, AufenthG).

and cooperation with BAMF. At the time of design of the present study in October 2015, permission was denied.

Although the AZR is the best sampling frame for a nationwide, multi-stage stratified random sample of asylum seekers (Schnell et al., 2013a, p. 260), it was only partially valid in 2015 and 2016 due to the rapid increase in the number of persons seeking asylum in Germany. Inaccuracies in the AZR resulted from the delayed registration of asylum seekers and from the use of two separate registration systems: (a) the AZR and (b) EASY, the then newly established system for the initial registration and distribution of asylum seekers among the German federal states on the basis of the above-mentioned quota system, the Königstein key. The use of two registration systems can give rise to duplicate records and to the underestimation of the number of cases (BAMF, 2015). Furthermore, it can be expected that a certain percentage of refugees will have left Germany without applying for asylum. The use of EASY data for research purposes is problematic because sex and age are not recorded. Asylum seekers nationwide are registered in the AZR when they file a formal application for asylum. The number of asylum seekers registered in the EASY system in 2015 (1,091,894) was more than twice as high as the number of first-time asylum applications filed in that year (441,899; BAMF, 2016b, p. 10). Whereas BAMF (2016b, p. 10) attributed this discrepancy to registration errors and double registrations in the EASY system, Kroh et al. (2017, p. 7) also drew attention to the so-called “EASY gap,” that is, the time lag between initial registration in the EASY system and the filing of a formal application for asylum, which was particularly pronounced in 2015. The introduction of a “proof of arrival” (Ankunftsnachweis, AKN) card for asylum seekers, as well as the integration of the registration systems of the individual authorities at federal, regional, and local level into a so-called “core data system,” enabled the number of asylum seekers who arrived in Germany in 2015 to be determined more precisely. As a result, the revised figure (890,000) was considerably lower than the previously assumed 1.1 million persons (BMI, 2016). By refreshing the sample several times, the IAB-BAMF-SOEP Survey of Refugees was able to overcome the discrepancies in the AZR (Kroh et al., 2017, p. 7).

For the reasons mentioned above, the AZR was not considered a suitable database for sampling. Hence, even if we had been given access to the AZR for this purpose, a representative sample could not have been drawn for Bavaria. However, the AZR was used in the present study as a data basis for quota sampling, as it enables the drawing of statistics on the number, nationality, and age and sex structure of asylum seekers aggregated at the level of foreigners authorities in cities and rural districts.

To sum up: As other methods of random sampling for rare populations were not applicable, a quota sampling design (Schnell et al., 2013a, p. 294) was used for the survey conducted within the framework of the present study.

Target Population and Sampling

The target population of the sample consisted of asylum seekers who arrived in Germany between January 2015 and February 2016 and who, in the light of high protection rates, had prospects of remaining in the country. At the end of 2015, when the pilot study was designed, this applied to asylum seekers from Syria (protection rate: 97.9%), Eritrea (92.2%), Iraq (70.2%), and Afghanistan (55.8%) (BAMF, 2017a, p. 50). In order to represent the differing circumstances among asylum seekers in Bavaria, legal asylum status was not a criterion for selection into the sample. Therefore, the target population comprised persons who had already applied for asylum and persons who had not yet had the chance to apply for asylum due to administrative delays.

According to AZR data, more than half of the asylum seekers who arrived in Bavaria between January 1 and September 30, 2015 were from Syria; 23% were from Afghanistan, 14% from Iraq, and 11% from Eritrea (see Table 1). These four countries were therefore the target countries of origin in the present study. At the time, the sex ratio among all asylum seekers in Bavaria irrespective of country of origin was 20 females to 80 males, yet the percentage of females among the asylum seekers from the aforementioned four target countries of origin was 25.6%.

While awaiting a decision on their asylum application, asylum seekers live in collective accommodation located in urban districts and rural communities. The research sites for the present study were selected with the aim of including heterogeneously structured areas and surveying a sufficient number of asylum seekers from the four target countries of origin, Syria, Iraq, Afghanistan, and Eritrea. The city of Nuremberg was chosen to represent urban areas and the district of Ebersberg, situated on the outskirts of Munich, to represent rural areas.

The target population included all residents aged 18 years or over living in collective accommodation for asylum seekers at the selected locations. The quota sample controlled for country of origin and sex (see Table 3). Although unit nonresponse was not reported systematically, the interpreters who delivered the questionnaires reported that almost all the target residents participated in the survey. However, in future studies, more exact measurement of nonresponse rates would be useful for further analysis.

Due to the quota sampling frame employed, the present study does not claim to be representative. However, as a pilot study, it provides an initial insight into the motives, attitudes, and intentions of asylum seekers. Due to controlled residence allocation – asylum seekers are supposed to be distributed among German federal states according to the Königstein key and irrespective of personal attributes – little selection bias was expected. It is therefore assumed that the motives and attitudes of newly arrived asylum seekers are not related to their current place of residence. Nevertheless, there may be bias: Some asylum seekers may have left the place of

Table 1 Sex ratio of asylum seekers from selected countries of origin in Bavaria (January–September 2015)

	Male		Female		Total	
Country of Origin	No.	%	No.	%	No.	%
Syria	6,041	82.4	1,290	17.6	7,341	51.8
Iraq	1,385	67.9	655	32.1	2,044	14.4
Afghanistan	2,687	81.5	609	18.5	3,306	23.3
Eritrea	1,150	77.7	331	22.3	1,487	10.5
Total	11,263	74.4	2,885	25.6	14,178	100

Source: AZR/BAMF, as of December 10, 2015; aggregate data on asylum seekers in Bavaria, January 1–September 30, 2015; own analysis of special analyses made available to the authors by BAMF.

Table 2 Population of asylum seekers at the research sites (as of April/May 2016)

	Nuremberg		Ebersberg		Total	
Country of Origin	No.	%	No.	%	No.	%
Syria	1,488	34.5	220	14.3	1,708	29.2
Iraq	1,162	27.0	61	4.0	1,223	20.9
Afghanistan	68	1.6	220	14.3	288	4.9
Eritrea	0	0.0	255	16.6	255	4.4
Other countries	1,591	36.9	784	50.9	2,375	40.6
Total	4,309	100	1,540	100	5,849	100

Source: (a) Fachstelle für Flüchtlinge (Specialist Unit for Refugees) at the Social Welfare Office of the City of Nuremberg (as of April 30, 2016). (b) Landratsamt Ebersberg (Ebersberg Administration Office; as of May 10, 2016).

residence allocated to them after their arrival and moved elsewhere (within Germany or abroad). As a consequence, asylum seekers in Bavaria might be different from those living in other federal states, for example Berlin or North Rhine-Westphalia. It has also been reported that asylum seekers sent to other federal states have sometimes returned to Bavaria. Moreover, it is possible that asylum seekers are allocated to collective accommodation based on their nationality. One indicator for this assumption is that there are only small numbers of asylum seekers from Afghanistan and Eritrea in Nuremberg (see Table 2).

Table 3 Quota plan by country of origin and sex

Country of Origin	Male	Female	Total
Syria	300	100	400
Iraq	190	60	250
Afghanistan	50	–	50
Eritrea	50	–	50
Total	590	160	750
Percent	78.7%	21.3%	100%

The survey design implemented was aimed at interviewing all asylum seekers from the target group allocated to the collective accommodation centers that were selected as research sites. As legal asylum status was not a selection criterion, persons who had already applied for asylum and persons who had not yet had the chance to apply for asylum due to administrative delays were eligible to participate in the survey.

As mentioned above, an analysis of the locally registered asylum seekers revealed that there was a relatively small number of persons from Afghanistan and Eritrea residing in Nuremberg (see Table 2). Moreover, in the rural district of Ebersberg near Munich, there were hardly any women among the asylum seekers from the target countries.

A separate evaluation of subpopulations, such as Eritrean or Afghan women, was not possible due to the small number of cases. It was therefore decided to limit the target female population to women from Syria and Iraq. A quota scheme based on country of origin and sex was designed for the sampling procedure (see Table 3). The study was scaled for at least 750 interviews, 20% of which were to be conducted with female respondents.

Standardized Survey Design and Data Collection

After conducting expert interviews with persons entrusted with the accommodation, distribution, and integration of asylum seekers, a standardized questionnaire was designed and translated into the most frequently used languages in the target population: English, Arabic, Farsi, and Tigrinya. When necessary, the interpreters translated the Arabic-language questionnaire into Kurdish (especially Kurmanji-Kurdish, the dialect spoken in Syria and northern Iraq), as Kurdish is rarely used as a written language. Based on existing literature (see Becher & El-Menouar, 2013;

Halm & Sauer, 2015; Pollack & Müller, 2013; Haug et al., 2009; Wetzels & Brettfeld, 2007) and the findings of the expert interviews, the following topics were included in the survey: place of origin and family structures; reasons for flight; intentions to remain in Germany; resources for structural integration (language skills, level of education, work experience); and perceptions and attitudes (toward gender roles, religiosity, tolerance, antisemitism, terrorism, democracy).

A self-administered paper-and-pencil (PAPI) mode was used to interview respondents. The advantage of a self-administered survey is that it reduces interviewer effects, such as response bias based on social desirability (Schnell et al., 2013a, p. 359). However, in an exclusively self-administered survey, the high rate of non-formal education among asylum seekers (Rich, 2016, p. 5), especially those from Afghanistan and Eritrea, would have led to bias due to educational background. Therefore, respondents with no literacy skills were able to complete the questionnaire with the help of native-speaker interpreters in a face-to-face interview setting.

A pretest with 10 respondents was conducted in a collective accommodation center for asylum seekers in Nuremberg in April 2016. After an initial evaluation of the results of the pretest, questions and items were modified. The pretest highlighted challenges relating to the organization of the field phase in collective accommodation for asylum seekers: The inclusion of “gatekeepers” was essential in order to gain access. Gatekeepers in relevant key positions were local government employees and employees of welfare organizations, as well as accommodation center managers and volunteers. Permission to access the collective accommodation had to be obtained from the agency responsible. The management of each accommodation was contacted and given information about the project and a description of the research design. Prior to the field work, information sheets in the relevant languages were posted in the collective accommodation in order to inform residents about the purpose and content of the study, as well as the survey period.

Eight native-speaker interpreters for Arabic/Kurdish, Farsi, and Tigrinya, who were also fluent in German, were hired to approach potential respondents. They informed them about the content of the study and data protection and anonymity issues, and they handed out the questionnaires in the relevant language. To prevent response bias, interpreters were specifically instructed to emphasize that participation was voluntary and independent of the legal asylum process. Addressing asylum seekers in their mother tongue proved to be a crucial element in establishing trust and ensuring participation in the survey. Interpreters were instructed to remain nearby while respondents completed the questionnaire, so that they could clarify any issues regarding the questions. All interpreters had an academic background; they had either completed a university degree in their home country or were currently studying at a university in Germany. In training sessions prior to the data collection process, interpreters were informed about the study and trained in the use of the research instruments and in appropriate behavior in collective accommodation centers for asylum seekers. Most interpreters had already been involved in the process of translating the questionnaires and were familiar with the content.

Fieldwork and Challenges

The fieldwork for the study was conducted in June and July 2016. Asylum seekers showed strong interest in participating in the study; this has also been the case in other surveys on asylum seekers (see, e.g., Brücker, Rother, & Schupp, 2016c). Participation incentives, such as pens, writing pads, and cloth bags, were offered to the respondents.

The illiteracy rate turned out to be only 5% of the entire sample, which meant that 95% of the respondents were able to self-administer the questionnaire. The interview duration ranged between 10 and 20 minutes. A total of 779 interviews were conducted with asylum seekers from Syria, Eritrea, Iraq, and Afghanistan at the two research locations, Nuremberg and Ebersberg (see Table 4).

As the quantitative study was designed as a self-administered survey, the influence of the interpreters, and therefore interviewer effects, were assumed to be marginal. This assumption was tested and proven.[2]

The analysis of the interpreters' field reports revealed that the respondents initially displayed a tendency to respond to questions about attitudes and values (e.g., concerning gender roles) in a socially desirable way (Paulhus, 2002, p. 50): During the interviews, they asked the interpreters for their opinions on appropriate answer patterns. The presence of the interpreters was important to explain to the respondents that their personal opinions were relevant and perfectly acceptable responses.

Some respondents tended to rely on country-specific response patterns instead of on their own opinions. Therefore, the cultural context must be taken into account when adapting survey instruments. Another option could be to give recently arrived respondents separate answer options regarding Germany and the country of origin in order to enable the expression of ambivalent attitudes in relation to the country of origin and the host country. For example, some respondents did not support the idea of women going out on their own at night. A deeper examination of this attitude item in the qualitative interviews revealed that this answer has to be interpreted in a political/cultural context: The security situation in countries of origin such as Iraq or Syria would not allow such behavior for safety reasons. Other respondents

[2] In the case of most of the attitudinal items, intraclass correlation (ICC) showed unremarkable values ranging from 2.6% to 6.3%. Some items showed higher ICCs; however, this was due to value discrepancies related to the country of origin: When controlled for country of origin, the interviewer effects proved to be marginal. Item nonresponse bias was a problem in the case of the item on antisemitism, as respondents from Eritrea had no concept of Jewish people (see Haug et al., 2017, p. 69).

Table 4 Sample by sex and country of origin

Country of Origin	Male	Female	Total
Syria	306	107	413
Iraq	190	62	252
Afghanistan	52	-	52
Eritrea	62	-	62
Total	610	169	779
Sex ratio	78.3%	21.7%	

Source: Dataset collected within the framework of the quantitative survey for the present study, *Asylsuchende in Bayern* (Asylum Seekers in Bavaria; Haug et al., 2017).

accepted this behavior on the part of women, but qualified their acceptance with reference to Germany; they argued that in their home countries this behavior would not be compatible with social norms and values.

Interpreters' reflections on the data collection process also indicated that religious and ethnic conflicts in the country of origin affected response behavior in the survey. For example, two respondents who self-identified as Sunni Muslims were clearly identified by the interpreters – on the basis of their hometown in Syria – as Alawis.

A further challenge arose from the interview setting in large collective accommodation centers for asylum seekers and, in particular, in emergency accommodation. Recruitment gained momentum when the presence of the interpreters attracted the interest of a number of residents: A whole group of asylum seekers could be informed about the survey at the same time, which significantly facilitated the recruitment and information process. On the other hand, conducting interviews in a collective residential setting can lead to socially desirable responses influenced by the proximity of other residents. The intervention of the interpreters was sometimes necessary to control these influencing tendencies by pointing out that the questionnaire was to be completed individually. Therefore, interpreters played an important role in avoiding bias due to socially desirable responding related to value concepts specific to the home country. The interpreters' explanations about how to complete a questionnaire were highly relevant because, for many respondents, it was the first survey of their lives.

Qualitative Methods

The qualitative study employed semi-structured face-to-face interviews with biographical elements. The interviews were conducted between June and October 2016 with the support of consecutive interpreters.

Qualitative research methods are used in cases where research topics cannot be investigated well using standardized methodological procedures and where an in-depth understanding of phenomena and their interpretation is required (Helfferich, 2011). The present study employed qualitative methods to explore attitudes and value orientations of asylum seekers, their reasons for flight, and their aspirations for the future. The study included (retrospective) biographical research techniques, which can be used to describe and analyze changes in (behavioral) patterns over time and cause–effect relationships (Fuchs-Heinritz, 2009; Rosenthal, 2004). The biographical approach allowed a contextual understanding of reasons for flight and other relevant dimensions in the genesis of biographical experiences. It was used to explain and interpret current phenomena, for example employment or training opportunities, and intentions to remain in Germany.

Participant Selection and the Use of Gatekeepers

The selection of 12 participants for the qualitative study was based on a theoretical sampling strategy (Marshall, 1996) that aimed to ensure diversity in the sociodemographic profiles of the interviewees with regard to country of origin, sex, age, and research location (see Table 5).

The use of gatekeepers (Creswell, 2003; Helfferich, 2011) in this study had a positive effect on obtaining access to asylum seekers in collective accommodation. The recruitment of research participants required official approval at the political and administrative levels, as well as the support of the accommodation providers, accommodation management, and security services on the ground. Gatekeepers, for example integration commissioners, were also helpful in finding potential interview candidates, coordinating and organizing the interviews, and recruiting asylum seekers according to the specific criteria of the sampling frame. In order to control for possible bias due to selection effects (e.g., high motivation to participate), an additional three participants were approached and recruited in situ at collective accommodation centers. As in other qualitative studies on asylum seekers (e.g., Brücker et al., 2016b), high motivation to participate in the qualitative interviews was observed.

Table 5 Cases in the qualitative survey

	Country of Birth	Sex	Age	Arrival in Bavaria	Asylum Application	Marital Status
1	Syria	M	18	2015	submitted	unmarried
2	Syria	M	21	2015	submitted	unmarried
3	Syria	M	37	2015	submitted	married, 3 children
4	Syria	F	27	2015	submitted	married, 4 children
5	Iraq	M	19	2016	submitted	unmarried
6	Iraq	M	27	2016	submitted	unmarried
7	Iraq	M	51	2015	submitted	unmarried
8	Iraq	F	32	2015	submitted	married, 3 children
9	Eritrea	M	19	2015	submitted	unmarried
10	Eritrea	M	41	2015	approved	married, 3 children
11	Afghanistan	M	22	2015	submitted	unmarried
12	Afghanistan	M	25	2015	submitted	unmarried

Qualitative Data Collection

All interviews were conducted with the support of consecutive interpreters by a 38-year-old male German researcher experienced in the field of qualitative biographical data collection. The interviews took place in collective accommodation for asylum seekers or at facilities of supporting local councils; interviews were audio-recorded.

At the beginning of each interview, the researcher informed the interviewee about the study and stressed that his or her personal data would be anonymized. The respondent was asked to sign a consent form, which had been drafted in his or her mother tongue. It was also emphasized that the study was not related to the interviewee's personal asylum procedure. The interviews lasted up to three and three quarter hours; a semi-structured interview guide was used, which focused on themes such as personal biography, experiences during flight, experiences in Germany, attitudes and value orientations, and future aspirations.

Visualization techniques were used to minimize memory effects on biographical data. A “life history guide” was developed, which facilitated structured biographical data collection (see Denzin & Lincoln, 1998). The life history guide is a paper-and-pencil technique that shows a simple time line on which biographical events are recorded. The time line follows the life course as it develops, and it allows the researcher to make structured notes as anchor points that can stimulate further memories of the respondent.

In addition, maps were given to the interviewees to enable them to locate geographically significant locations in their home country and on the route they took when they fled. Reporting the individual way stations proved difficult for most participants, which suggests that orientation on the route could depend more on external factors and actors, such as traffickers or fellow refugees.

The Role of Interpreters in the Research Process

The literature suggests that interpreters play an important role in the qualitative interview process with asylum seekers (see, e.g., Brücker et al., 2016a; Brücker et al., 2016b). Interpreters were trained prior to data collection, and it was advantageous that they had already been involved in the translation of the semi-structured questionnaire and were familiar with the content of the study.

Ten interviews were conducted in Tigrinya or Arabic with male interpreters; two interviews were conducted in Farsi with a female interpreter. All interpreters were native speakers of the respective languages and were also fluent in German. The fact that the interpreters and asylum seekers had a common cultural and language background helped build interviewees’ confidence in the study project and make the interviews more enjoyable and effective. Moreover, the fact that some of the interpreters originally came to Germany as asylum seekers themselves had an additional confidence-building effect.

The use of interpreters can influence interviews (the so-called “interpreter effect”; see Jentsch, 1998). Interpreters can affect response behavior due to their presence, behavior, and external characteristics (e.g., sex and age) and thus create bias. Negative effects on the interview process due to different cultural/religious backgrounds on the part of interpreters and respondents (e.g., Kurds, Shia Arabs, & Sunni Arabs) could not be observed in this study (but see, e.g., Jacobsen & Landau, 2003).

Two Arabic-speaking female asylum seekers were interviewed by a male interviewer with the support of a male interpreter. In one case, the husband was present at the interview but did not interfere with the interview process. A female also acted as interpreter at interviews with male asylum seekers. In both constellations, no significant gender effects in the form of refusal or response bias could be determined.

Another issue was the discrepancy between literal and free translation, which can significantly influence the interview situation. The present study used an approach that gave the interpreter the option to translate interview questions and answers largely freely. This was also helpful because differing cultural backgrounds, as well as low levels of education on the part of interviewees, sometimes made it necessary to provide explanations for certain terms, such as "integration".

In some cases, native-speaker transcribers were hired to check the translations of interpreters during the interviews (see Merkens, 1997). In these cases, the Arabic content of a recorded interview was additionally translated by a native-speaker transcriber. The comparison of the Arabic originals with the control translations did not reveal any relevant differences. The transcriptions of the other interviews focused only on the German text.

Research Ethics

The principles for research ethics drawn up by the German Data Forum (RatSWD, 2017) focus in particular on interviews with vulnerable groups. Studies on asylum seekers present specific research-ethical challenges concerning research design and the process of data collection (Hugman et al., 2011; Jacobsen & Landau, 2013), a fact that played an important role in the present study.

All persons involved in this study (including researchers, interviewers, interpreters, and transcribers) were obliged to maintain strict confidentiality of the information obtained in the course of the study. Survey respondents and qualitative interviewees were given written information on the content and objectives of the interview and on data protection, as well as a guarantee of confidentiality of interview contents and personal data. The explanatory note on the quantitative survey pointed out the study's voluntary nature and anonymity, as well as the fact that participation in the study would have no effect on the asylum procedure. For the qualitative study, a declaration of consent form was prepared and translated into the relevant languages. The declaration of consent form mentioned that participation in the study was voluntary, and expressly emphasized that the survey was unrelated to any possible asylum procedure. All qualitative interviews were subject to written consent to participation in the study. Depending on the educational background of the interviewee, an extensive explanation of the concept of declaration of consent was necessary.

To ensure interviewee anonymity in the evaluation process, the names of the persons were changed, and other data (e.g. places and chronological time lines) were represented so imprecisely in the results that re-identification is not possible. Here, the blurring of the data material presented was consciously weighed up against the requirement to protect the privacy of the respondents. The original data material (audio files, transcripts) was stored securely and will be deleted after completion of the project. The anonymized transcripts were stored and handed over to Hanns Seidel Foundation as an appendix to the final report. Whereas the final report was published (Haug et al., 2017), this appendix was not.

As asylum seekers are a vulnerable group of research subjects, difficult interaction situations may arise during the research process (Helfferich, 2011). Both traumatic events in the country of origin and experiences when fleeing to Germany may cause post-traumatic stress disorder (PTSD), and this should be taken into account during biographical interviews, which may cause interviewees to relive traumatic events. None of the respondents showed obvious negative emotional responses to the interviews or exhibited signs of an emotional crisis. In order to fulfill their ethical responsibility, the researchers made sure that information on psychotherapeutic care facilities was passed on to the participants in the course of the study.

Conclusion

In many respects, research on asylum seekers is not easy to conduct. As a pilot study, the research project “Asylum Seekers in Bavaria” offered the opportunity to test methodological approaches to conducting a quantitative and qualitative survey of persons who had recently arrived in Germany in search of asylum. In particular, the use of gatekeepers and interpreters proved to be an essential feature of the research process. The cooperation of native speakers and transcribers was also essential for the interpretation of the data. In order to build trust among the potential respondents, it was important to reduce uncertainties by explaining the rules and concepts of data protection and anonymity and to point out the strict separation of the present research from the asylum procedure. The majority of asylum seekers greatly appreciated being approached in their native language; this was reflected in a high willingness to participate.

One challenge was the sampling of asylum seekers in the quantitative survey. Due to the strong influx of asylum seekers between autumn 2015 and spring 2016, there was no sufficiently valid database at time of the project design that included all asylum seekers who entered Bavaria. The AZR database lists all asylum seekers in Germany, but, as explained above, the deviations between the AZR and the EASY registration database (the so-called “EASY gap”; Kroh et al., 2017, p. 7) were particularly pronounced in 2015. Furthermore, due to statutory restrictions, samples can be drawn from the AZR only within the framework of research projects conducted in cooperation with BAMF, which was not the case in the present study; hence a pilot survey was conducted in two research areas in Bavaria.

The use of qualitative interviews in a mixed-methods approach proved helpful, in particular, to interpret the response patterns of research participants in the quantitative survey. For example, findings from the qualitative interviews suggest that it was problematic for interviewees to answer questions on attitudes to religiosity, freedom of opinion, and gender roles, which were also asked in the quantitative survey.

In 2017, a second wave of qualitative interviews took place in order to collect exemplary integration biographies (Haug & Huber 2018). However, as the 2016 quantitative study was designed as a cross-sectional study, no statements can be made about general integration processes or the determinants of integration based on the presented case study. The IAB-BAMF-SOEP Survey of Refugees, a German Socio-Economic Panel (GSOEP) study, will provide such data for future research.

References

Becher, I., & El-Menouar, Y. (2013). *Geschlechterrollen bei Deutschen und Zuwanderern christlicher und muslimischer Religionszugehörigkeit* [*Gender roles among Germans and immigrants of Christian and Muslim faith*] (BAMF Research Report No. 21). Nuremberg: Federal Office for Migration and Refugees (BAMF).

Brücker, H., Fendel, T., Kunert, A., Mangold, U., Siegert, M., & Schupp, J. (2016b). *Geflüchtete Menschen in Deutschland: Warum sie kommen, was sie mitbringen und welche Erfahrungen sie machen* [*Refugees in Germany: Why they come, what they have to offer, and what they experience*] (IAB Short Report No. 15/2016). Nuremberg: Institute for Employment Research (IAB).

Brücker, H., Kunert, A., Mangold, U., Kalusche, B., Siegert, M., & Schupp, J. (2016a). *Geflüchtete Menschen in Deutschland – eine qualitative Befragung* [*Refugees in Germany – A qualitative survey*] (IAB Research Report No. 9/2016). Nuremberg: Institute for Employment Research (IAB).

Brücker, H., Rother, N.; & Schupp, J. (Eds.). (2016c). *Die IAB-BAMF-SOEP-Befragung von Geflüchteten: Überblick und erste Ergebnisse* [*The IAB-BAMF-SOEP Survey of Refugees: Overview and first results*]. Berlin: German Institute for Economic Research (DIW).

BAMF (Federal Office for Migration and Refugees). (2015). *Sehr hoher Asylzugang im September* [*Very high number of asylum seekers in September*] (press release issued October 7, 2015). Retrieved December 15, 2016 from http://www.bamf.de/SharedDocs/Meldungen/DE/2015/20151007-asylgeschaeftsstatistik-september.html

BAMF (Federal Office for Migration and Refugees). (2016a). *Migrationsbericht 2015* [*Migration report 2015*]. Nuremberg: BAMF.

BAMF (Federal Office for Migration and Refugees). (2016b). *Das Bundesamt in Zahlen 2015* [*The Federal Office in figures 2015*]. Nuremberg: BAMF. Retrieved September 7, 2018 from https://www.bamf.de/SharedDocs/Anlagen/DE/Publikationen/Broschueren/bundesamt-in-zahlen-2015.pdf

BAMF (Federal Office for Migration and Refugees). (2017a). *Das Bundesamt in Zahlen 2016* [*The Federal Office in figures 2016*]. Nuremberg: BAMF. Retrieved September 7, 2018 from http://www.bamf.de/SharedDocs/Anlagen/DE/Publikationen/Broschueren/bundesamt-in-zahlen-2016.pdf

BAMF (Federal Office for Migration and Refugees). (2017b). *Erstverteilung der Asylsuchenden (EASY)* [*Initial Distribution of Asylum Seekers (EASY)*] (as of January 1, 2017). Retrieved June 23, 2017 from http://www.bamf.de/DE/Fluechtlingsschutz/AblaufAsylv/Erstverteilung/erstverteilung-node.html

BAMF (Federal Office for Migration and Refugees) (Eds.). (2018) *Aktuelle Zahlen zu Asyl, Juli 2018* [*Current figures on asylum, July 2018*] (monthly report). Retrieved September 7, 2018 from http://www.bamf.de/SharedDocs/Anlagen/DE/Downloads/Infothek/Statistik/Asyl/aktuelle-zahlen-zu-asyl-juli-2018.pdf

BMI (German Federal Ministry of the Interior). (2016). *890.000 Asylsuchende im Jahr 2015* [*890,000 asylum seekers in 2015*] (press release issued September 30, 2016). Retrieved October 14, 2016 from https://www.bmi.bund.de/SharedDocs/Pressemitteilungen/DE/2016/09/asylsuchende-2015.html

Creswell, J. (2003). *Research design*. Thousand Oaks, CA: Sage Publications.

Denzin, N., & Lincoln, Y. (1998). Entering the field of qualitative research. In N. Denzin & Y. Lincoln (Eds.), *Strategies of Qualitative Inquiry* (pp. 1–34). Thousand Oaks, CA: Sage Publications.

Fuchs-Heinritz, W. (2009). *Biographische Forschung* [*Biographical research*]. Wiesbaden: VS Verlag.

Halm, D., & Sauer, M. (2015): *Lebenswelten deutscher Muslime* [Lifeworlds of German Muslims] (research report in the Bertelsmann Foundation series *Religionsmonitor*). Gütersloh: Bertelsmann.

Haug, S., & Huber, D. (2018). *Asylsuchende in Bayern. Eine qualitative Folgebefragung.* München: Hanns-Seidel-Stiftung.

Haug, S., Currle, E., Lochner, S., Huber, D., & Altenbuchner, A. (2017). *Asylsuchende in Bayern* [*Asylum seekers in Bavaria*] (research report). Munich: Hanns Seidel Foundation. Retrieved June 23, 2017 from https://www.hss.de/download/publications/Asylsuchende_in_Bayern.pdf

Haug, S., & Vernim, M. (2015). *Telefonische Befragung. Methodenbericht. Der Einfluss sozialer Netzwerke auf den Wissenstransfer am Beispiel der Reproduktionsmedizin (NeWiRe)* [*Telephone survey. Methods report. The influence of social networks on knowledge transfer, using the example of reproductive medicine*] (Working Paper No. 2.01). Regensburg: OTH. Retrieved January 3, 2018 from https://www.oth-regensburg.de/fileadmin/media/fakultaeten/s/forschung_projekte/IST/newire/NeWiRe_2.01_Methodenbericht_Telefonbefragung.pdf

Haug, S., Vernim, M., Gelfert, V., & Reindl, A. (2014). *Integrationsbericht und Integrationskonzept für Regensburg* [*Integration report and integration concept for Regensburg*] (final report). Regensburg: City of Regensburg/OTH Regensburg.

Haug, S., Müssig S., & Stichs, A. (2009). *Muslim life in Germany* (Research Report No. 6). Nuremberg: Federal Office for Migration and Refugees (BAMF). Retrieved October 19, 2018 from https://www.bamf.de/SharedDocs/Anlagen/EN/Publikationen/Forschungsberichte/fb06-muslimisches-leben.pdf?__blob=publicationFile

Helfferich, C. (2011). *Die Qualität qualitativer Daten. Manual für die Durchführung qualitativer Interviews* [*The quality of qualitative data. A manual for conducting qualitative interviews*]. Wiesbaden: VS Verlag.

Hugman, R., Pittaway, E., & Bartolomei, L. (2011). When ‘do no harm’ is not enough: The ethics of research with refugees and other vulnerable groups. *British Journal of Social Work, 41*(7), 1271–1287.

Humpert, A., & Schneiderheinze, K. (2002) Stichprobenziehung für telefonische Zuwandererbefragungen – Erfahrungen und neue Ansätze [Drawing samples for telephone surveys of immigrants – Experiences and new approaches]. In S. Gabler & S. Häder (Eds.), *Telefonstichproben – Methodische Innovationen und Anwendungen in Deutschland* [*Telephone Samples – Methodological Innovations and Applications in Germany*] (pp 187–208). Münster: Waxmann.

Jacobsen, K., & Landau, L. (2003). The dual imperative in refugee research: Some methodological and ethical considerations in social science research on forced migration. *Disasters*, *27*(3), 185-206.

Jentsch, B. (1998). The 'interpreter effect': Rendering interpreters visible in cross-cultural research and methodology. *Journal of European Social Policy*, *8*(4), 275–289.

Johannsson, S. (2016). *Was wir über Flüchtlinge (nicht) wissen. Der wissenschaftliche Erkenntnisstand zur Lebenssituation von Flüchtlingen in Deutschland* [*What we (don't) know about the living conditions of refugees in Germany*] (expert report on behalf of the Robert Bosch Foundation and the Expert Council of German Foundations on Integration and Migration, SVR). Berlin: SVR. Retrieved August 11, 2016 from http://www.svr-migration.de/wp-content/uploads/2016/01/Was-wir-%C3%BCber-Fl%C3%BCchtlinge-nicht-wissen.pdf

Kroh, M., Kühne, S., Jacobsen, J., Siegert, M., & Siegers, R. (2017). *Sampling, nonresponse, and integrated weighting of the 2016 IAB-BAMF-SOEP Survey of Refugees (M3/M4) – Revised version* (SOEP Survey Paper No. 477: Series C). Berlin: DIW/SOEP.

Liebau, E., Humpert, A., Schneiderheinze, K. (2018). *Wie gut funktioniert das Onomastik-Verfahren? Ein Test am Beispiel des SOEP-Datensatzes* [*How well does the onomastic method work? A test using the example of the SOEP dataset*] (SOEP Survey Paper No. 976). Berlin: DIW/ SOEP.

Marshall, M. (1996). Sampling for qualitative research. *Family Practice*, *13*, 522-526.

Merkens, H. (1997). Stichproben bei qualitativen Studien [Sampling in qualitative studies]. In B. Friebertshäuser & A. Prengel, (Eds.), *Handbuch Qualitative Forschungsmethoden in der Erziehungswissenschaft* [*Handbook of Qualitative Research Methods in Educational Science*] (pp. 97–106). Munich: Juventa.

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

Pollack, D., & Müller, O. (2013). *Verstehen was verbindet. Religiosität und Zusammenhalt in Deutschland* [*Understanding what unites people. Religiosity and cohesion in Germany*] (research report in the Bertelsmann Foundation *Religionsmonitor* series). Gütersloh: Bertelsmann.

RatSWD (German Data Forum). (2017). *Forschungsethische Grundsätze und Prüfverfahren in den Sozial- und Wirtschaftswissenschaften* [*Principles and review procedures for research ethics in the social and economic sciences*] (Output 9 (5)). Berlin: RatSWD.

Rich, A. (2016). *Asylantragsteller in Deutschland im Jahr 2015. Sozialstruktur, Qualifikationsniveau und Berufstätigkeit* [*First-time asylum applicants in Germany in 2015. Social structure, level of qualifications and employment*] (BAMF Brief Analysis No. 3/2016). Nuremberg: Federal Office for Migration and Refugees (BAMF). Retrieved January 1, 2017 from http://www.bamf.de/SharedDocs/Anlagen/EN/Publikationen/Kurzanalysen/kurzanalyse3_sozial-komponenten.pdf?__blob=publicationFile

Rosenthal, G. (2004). Biographical research. In C. Seale, G. Gobo, J. Gubrium, & D. Silverman (Eds.). *Qualitative Research Practice* (pp. 48–64). London: Sage Publications.

Schnell, R., Gramlich, T., Bachteler, T., Reiher, J., Trappmann, M., Smid, M., & Becher, I. (2013b). Ein neues Verfahren für namensbasierte Zufallsstichproben von Migranten [A new name-based sampling method for migrants]. *mda Methoden - Daten – Analysen*, *7*(1), 5–33.

Schnell, R., Hill, P., & Esser, E. (2013a). *Methoden der empirischen Sozialforschung* [*Methods of empirical social research*] (10th. ed.). Munich: Oldenbourg.

VDSt Arbeitsgemeinschaft Bevölkerung (Association of German Municipal Statisticians Population Working Group). (2013). *Migrationshintergrund in der Statistik – Definitionen, Erfassung und Vergleichbarkeit* [*Migration background in statistics – Definitions, measurement, and comparability*] (issue 2 of the VDSt Population Working Group publications *Materialien zur Bevölkerungsstatistik*). Cologne: Association of German Municipal Statisticians (VDSt). Retrieved September 7, 2018 from http://www.staedtestatistik.de/fileadmin/vdst/AG_Bevoelkerung/Publikation/Heft2_Migrationshintergrund.pdf

Wetzels, P., & Brettfeld, K. (2007). *Muslime in Deutschland. Integration, Integrationsbarrieren, Religion, und Einstellungen zu Demokratie, Rechtsstaat und politisch-religiös motivierter Gewalt* [*Muslims in Germany. Integration, integration barriers, religion, and attitudes to democracy, the rule of law, and politically and religiously motivated violence*] (report of the results of a survey study). Berlin: Federal Ministry of the Interior (BMI).

Worbs, S., Bund, E., & Böhm, A. (2016). *Asyl – und dann? Die Lebenssituation von Asylberechtigten und anerkannten Flüchtlingen in Deutschland. BAMF-Flüchtlingsstudie 2014* [*Asylum – and then? The living conditions of persons granted asylum status and recognized refugees in Germany. BAMF Refugee Study 2014*] (Research Report No. 28). Nuremberg: Federal Office for Migration and Refugees (BAMF).

Thank to Reviewers

The Editors of methods, data, analyses would like to thank the following referees who have reviewed manuscripts for the journal from January 2018 to January 2019:

Christopher Antoun, Maryland
Sharon Baute, Leuven
Constanze Beierlein, Hamm
Michael Blohm, Mannheim
Michael Braun, Mannheim
Nate Breznau, Mannheim
Sarah Butt, London
Mario Callegaro, London
Jan Cieciuch, Warsaw
Eldad Davidov, Cologne
Brita Dorer, Mannheim
Hermann Duelmer, Cologne
Martin Elff, Friedrichshafen
Anke Erdmann, Bielefeld
Marieke Haan, Groningen
Dominique Joye, Lausanne
Markus Klein, Glasgow
Simon Kühne, Bielefeld
Heinz Leitgöb, Eichstaett-Ingolstadt
Noah Lewin-Epstein, Tel-Aviv
Jochen Mayerl, Chemnitz
Bart Meuleman, Leuven
Christian Monseur, Liège
Guy Moors, Tilburg
Cornelia Neuert, Mannheim
Frans Oort, Utrecht
Jost Reinecke, Bielefeld
Melanie Revilla, Barcelona
Angelika Scheuer, Mannheim
Elmar Schlüter, Giessen
Peter Schmidt, Giessen
Evi Scholz, Mannheim
Matthias Schonlau, Waterloo
Daniel Seddig, Cologne
Vera Toepoel, Utrecht
Klaus Troitzsch, Koblenz
Hagen von Hermanni, Leipzig
Brady Thomas West, Ann Arbor
Diana Zavala-Rojas, Barcelona
Conrad Ziller, Cologne

Information for Authors

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

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

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

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

- Please submit your manuscript via www.mda.gesis.org.
- The total length of the manuscript shall not exceed 10.000 words.
- Manuscripts should…
  - be written in English, using American English spelling. Please use correct grammar and punctuation. Non-native English speakers should consider a professional language editing prior to publication.
  - be typed in a 12 pt Roman font, double-spaced throughout.
  - be submitted 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.
  - should be anonymized ("blinded") for review.
- Please also send us an abstract of your paper (approx. 300 words), a front page with a brief biographical note (no longer than 250 words as supplementary file), and a list of 5-7 keywords for your paper.
- Acceptable formats for Graphics are
  - pdf
  - jpg (uncompressed, high quality)
- 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.
- Please number the headings of your article. Doing so will make the work of the layout editor easier.

- If your text includes formulas we would like to ask you to upload your text also as a PDF, additional to the Word document.
- By submitting a paper to mda the authors agree to make data and program routines available for purposes of replication.
- Response rates in research papers or research notes, where population surveys are analyzed, should be calculated according to AAPOR standard definitions.
- Please follow the APA guideline when structuring and formating your work.

When preparing in-text references and the list of references please also follow the APA guidelines:

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)