mda
methods, data, analyses
Journal for Quantitative Methods and Survey Methodology
Volume 10, 2016 | 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.gesis.org/mda

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

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

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

Print: Bonifatius Druck GmbH Paderborn, Germany

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

Content

RESEARCH REPORTS

methods, data, analyses | Vol. 10(2), 2016, pp. 119-142 DOI: 10.12758/mda.2016.011

Assessing the Use of Mode Preference as a Covariate for the Estimation of Measurement Effects between Modes. A Sequential Mixed Mode Experiment

Caroline Vandenplas[1], *Geert Loosveldt*[1] & *Jorre T. A. Vannieuwenhuyze*[2]

1 KU Leuven

2 Utrecht University

Abstract

Mixed mode surveys are presented as a solution to increasing survey costs and decreasing response rates. The disadvantage of such designs is the lack of control over mode effects and the interaction between selection and measurement effects. In a mixed mode survey, measurement effects can put into doubt data comparability between subgroups, or similarly between waves or rounds of a survey conducted using different modes. To understand the extent of measurement effects, selection and measurement effects between modes have to be disentangled. Almost all techniques to separate these effects depend on covariates that are assumed to be mode-insensitive and to fully explain selection effects. Most of the time, these covariates are sociodemographic variables that might be mode-insensitive, but fail to sufficiently explain selection effects. The aim of this research is to assess the performance of mode preference variables as covariates to evaluate selection and measurement effects between modes.

In 2012, a mixed mode survey – a web questionnaire followed by face-to-face interviews– was conducted alongside the face-to-face European Social Survey in Estonia (Ainsaar et al., 2013). The questionnaire included mode preference items. In this paper, the effects of the trade-offs between the two assumptions on the precision of estimated selection and measurement effects are compared. The results show that while adding the mode preference to the propensity score model seems to increase the explanatory power of web participation, it decreases the correlation between propensity scores and target variables. In addition, the estimated selection and measurement effects do not always fit the expectation that more selection effects are explained and more measurement effects are detected.

Keywords: mixed mode surveys; selection effects; measurement effects; mode preference; back-door method

1 Introduction

Mixed mode surveys are those during which different modes are offered simultaneously or sequentially. Such surveys have increased in popularity and are often implemented to adapt surveys to the needs or preferences of respondents. The implementation of mixed mode surveys is aimed at reducing costs, increasing response rates, and decreasing nonresponse bias, compared with traditional single mode surveys–especially face-to-face and telephone surveys. However, data collected using different modes may lead to differences in survey estimates due to mode effects. Mode effects can be separated into (1) selection effects, which are defined as differences in the responding sample due to different non-coverage or nonresponse errors between the modes, and (2) measurement effects, which occur when the answer from the same respondent would differ if a different data collection mode was used (Voogt & Saris, 2005; Weisberg, 2005).

1.1 Selection and Measurement Effects Between Modes

Selection effects between modes in a mixed mode survey can be desirable if they help to diversify the sample of respondents. Indeed, different modes may have different coverage problems and different levels of nonresponse bias (Dillman, Smyth, & Christian, 2009; de Leeuw, 2005). For example, the declining coverage of landline telephone surveys could be compensated for by adding a web questionnaire or face-to-face interviews for 'mobile only' individuals. Moreover, depending on their abilities and availability, individuals may be more likely to answer in one mode than in another. For example, web respondents are typically more likely to be higher educated and have a higher income, and are less likely to be elderly or from a minority compared with the general population. Indeed, people with these characteristics are more likely to be connected to the Internet, to use it frequently, and to have greater computer skills (Zillien & Hargittai, 2009; de Leeuw, 2005; Bimber, 2000). However, results concerning the benefits of using mixed mode surveys to reduce selection bias are mixed (e.g., Revilla, 2015; Medway & Fulton, 2012; Millar & Dillman, 2011; Holmberg et al., 2010; Smyth et al., 2010; US Census

Acknowledgements

We would like to thank the ESS Core Scientific Team for giving us access to the data, and the Estonian ESS team for the implementation of the European Social Survey mixed mode experiment in their country. We are also very grateful for the comments of two anonymous reviewers, which helped to improve the paper.

Direct correspondence to

Caroline Vandenplas, Center for Sociological Research, KU Leuven, Parkstraat, 45, 3000 Leuven, Belgium

E-mail: caroline.vandenplas@kuleuven.be

Bureau, 2010; Eva et al., 2010; Dillman, Phelps, et al., 2009; Gentry & Good, 2008; Fowler et al., 2002).

Measurement effects between modes may be problematic, especially if survey results need to be compared across rounds, across countries, or between subgroups in a country. When considering measurement effects, the measurement in one mode is often taken as the benchmark. Dillman (2000: chapter 6) points to differences in normative and cognitive consideration between modes, as well as interactions between the two. Especially when mixing interviewer-based and self-administered modes, the presence or absence of an interviewer and the aural or visual presentation of the items may lead to different stimuli and answering processes. The presence of an interviewer may increase socially desirable effects (the respondent taking social norms into consideration when answering the questions) and acquiescence (the tendency of the respondent to agree with the underlying statement of the question). Moreover, the visual presentation in a self-administered survey mode may increase primacy effects–choosing the first acceptable answer read–compared with aural presentation, which can favor recency effects–choosing the last acceptable answer heard. These effects can be reinforced by the lack of control over the cognitive efforts made by respondents in self-administered surveys, allowing them to not read the question and the answer options fully.

1.2 Back-door Method

Because the measured difference between alternative modes is a combination of selection and measurement effects, an important and complex issue is that of separating the two types of effects. To solve this confounding problem, Vannieuwenhuyze and colleagues (2010) suggest applying causal inference theory. In particular, the *back-door method* (Pearl, 2009; Morgan & Winship, 2009) can be applied to disentangle measurement and selection effects. The back-door method involves the inclusion of a set of variables X into the analysis model, where X explains the selection effects between different modes. The back-door method is based on two assumptions, the *mode selection ignorability assumption,* which requires that X fully captures the selection effects between the modes, and the *mode-insensitivity assumption*, which requires that the measurement of X is independent of the mode in which it is measured. Another proposed method to separate selection and measurement effects between modes is to re-interview respondents using another mode to estimate the measurement effects (Klausch, Hox, & Schouten, 2015; Klausch, Schouten, & Hox, 2015; Schouten et al., 2013).

Many existing attempts to separate selection effects from measurement effects in mixed mode surveys rely on the back-door method (e.g., Kolenikov & Kennedy, 2014; Vannieuwenhuyze et al., 2014; Vannieuwenhuyze & Loosveldt, 2013; Vannieuwenhuyze et al., 2012; Lugtig et al., 2011; Heerwegh & Loosveldt, 2011; Jäckle

et al., 2010; Hayashi, 2007). However, most of these attempts are based on a set of sociodemographic variables that can be argued to be mode-insensitive, but probably fail to fully explain selection effects, i.e. to make the groups responding in different modes comparable. Therefore, variables that can complement sociodemographic variables as covariates in the back-door method should be found, of which one example may be mode preference variables.

1.3 Mode Preference

Mode preference reflects the fact that there may be different modes (Groves & Kahn, 1979) in which sampled people are more likely to answer (Olson et al., 2012; Shi & Fan, 2007; Miller et al., 2002). Based on this preference, mixed mode surveys are expected to have better response rates, because the choice of data collection mode theoretically increases the response propensity. For instance, some people feel uncomfortable with web questionnaires, because they are not familiar with using computers or the Internet, whereas others may perceive a web questionnaire as less intrusive than a face-to-face interview (Smyth et al., 2014). Mode preferences can therefore be hypothesized to be good predictors of the selected survey mode in a mixed mode survey (Olson et al., 2012) and can act as back-door variables. However, questions about mode preference may be subject to measurement effects. Previous research shows that respondents are more likely to endorse the mode they participate in, and therefore in which the mode preference is measured (Millar, O'Neil, & Dillman, 2009; Gesell, Drain, & Sullivan, 2007; Tarnai & Paxson, 2004; Groves & Kahn, 1979).

Although such variables are not expected to fulfill the mode-insensitivity assumption, they may offer a better trade-off between compliance with the mode-insensitivity assumption and compliance with the mode-selection ignorability assumption, compared with using sociodemographic variables when evaluating measurement and selection effects between modes in a mixed mode survey. Moreover, a possible solution to this mode-sensitivity is the creation of a latent variable that allows the control of measurement effects between modes, using a multi-group structural model. This requires, of course, at least three items measuring mode preference.

1.4 Different Sets of Covariates for the Back-door Method

To test the hypothesis that mode preference variables achieve a better balance between the two assumptions, we compare three sets of variables in this article: Only sociodemographic, sociodemographic combined with mode preferences, and sociodemographic combined with a latent mode preference variable. On the one hand, selection effects could be underestimated when only sociodemographic vari-

ables are included as back-door variables. As a consequence, the selection effects would not be completely corrected when applying the back-door method and the residual selection effects would be wrongly attributed to the measurement effect. The measurement effects estimates would then be biased: Over or under-estimated if the selection and measurement effects are in respectively the same or the opposite direction. On the other hand, selection effects might be estimated more accurately when variables about mode preferences are included, given the expected strong relationship between mode selection effects and the mode preference variables. However, the consequences of the mode-sensitive nature of mode preference variables on the estimated selection effects are difficult to predict. They could accentuate the selection effects and lead to an overcorrection of the selection effect when applying the back-door method. Conversely, the mode-sensitivity of the mode preference could result in an underestimation of the selection effects, or even introduce a completely random component. Lastly, the inclusion as a covariate of a latent mode preference variable built on three measurements of mode preferences should allow for a more-precise estimation of the selection effects. Indeed, the latent variable is independent of random measurement errors on the three specific measurements, and forcing the structural model to be the same in both modes should reduce measurement effects.

2 Data

The European Social Survey (ESS) is an academically-driven survey, designed to study the interactions between changing institutions, attitudes, beliefs, and behavioral patterns in Europe. The ESS started in 2002 and has been repeated every two years. Since its first round, great efforts have been made within the ESS to collect high quality data, and to ensure cross-national and cross-cultural comparability. Given the issues of the increasing costs of face-to-face surveys and declining response rates in some countries, it was decided to explore the possibility of mixed mode survey designs as an alternative to the traditional face-to-face interviews.

In 2012, a mixed mode survey was conducted in Estonia in parallel to round six of the main ESS survey. A simple random sample of 925 individuals was drawn from the population register to participate in a sequential, mixed mode survey, involving a web questionnaire (mode *a*), followed by a face-to-face phase (mode *b*) for the sample units who did not participate in the web component. A first invitation letter to the web survey containing a hyperlink and an individual password was sent to the 925 sampled individuals on 18 September. Two reminders (copies of the invitation letter) were sent respectively two weeks and four weeks after the first invitation letter was sent, as well as a last reminder to people who started the online questionnaire without completing it within approximatively five weeks. On

22 October, the face-to-face stage started for all the sample units who had not completed the web questionnaire. In the end, 356 people (38.4%) responded via the web survey and 230 (24.8%) completed the face-to-face interview, making a total of 586 respondents. The final response rate of 63.3% is not significantly different from the response rate for the main ESS survey (2380 out of 3702 = 64.2%, Chi square p = 0.7), where response rates are calculated as the number of completed interviews/ questionnaires divided by the sample size, ignoring ineligible people.

An analysis of characteristics reveals some differences between the web and face-to-face respondents in the mixed mode survey. Results show that web respondents on average were younger, higher educated, and more likely to live in the North of Estonia compared with the face-to-face respondents (Ainsaar et al., 2013).

In addition to the usual ESS questionnaire, the mixed mode survey included questions about mode preference, survey attitudes, and the perceived accuracy of the survey.

The questionnaire contains three mode preference related variables that are considered as possible auxiliary variables to control for selection effects between the web and the face-to-face component of the survey. These variables are:

- Web participation (RPWEB): In general, how often would you respond to surveys like this one if you were invited to complete an internet questionnaire?
- Phone participation (RPPHONE): In general, how often would you respond to surveys like this one if you were invited to complete a telephone interview?
- Face-to-face participation (RPF2F): In general, how often would you respond to surveys like this one if you were invited to complete a face-to-face interview?

The answer categories are: 1= never, 2 = once in a while, 3 = about half of the time, 4 = most of the time, 5 = always. In the hope of reducing measurement effects between the modes, the variables related to mode preferences did not directly ask about the preferred mode, but were instead designed so that the mode preference could be deduced from them.

Item nonresponse to mode preference variables reduced the responding sample from 582 to 556. As a consequence, all analyses are performed considering these 556 respondents.

Table 1 displays the means and standard deviations of the three mode-preference variables among web respondents and among face-to-face respondents. The mean for 'phone participation' is similar between the two groups but the means for 'web participation' or 'face-to-face participation' are very different. As expected, web respondents have a higher mean for 'web participation' than face-to-face respondents, and face-to-face respondents have a higher mean for 'face-to-face participation' than web respondents.

Given the categorical nature of the variables, we also show the distribution of these variables in Figure 1.

Table 1 Means and standard deviations of the variables about mode preferences for web respondents and for face-to-face respondents

Variables	Web mean	Web standard deviation	Face-to-face mean	Face-to-face standard deviation
Web	2.50	1.19	1.60	0.97
Phone	1.56	0.95	1.71	1.06
Face-to-face	1.71	1.06	2.90	1.28

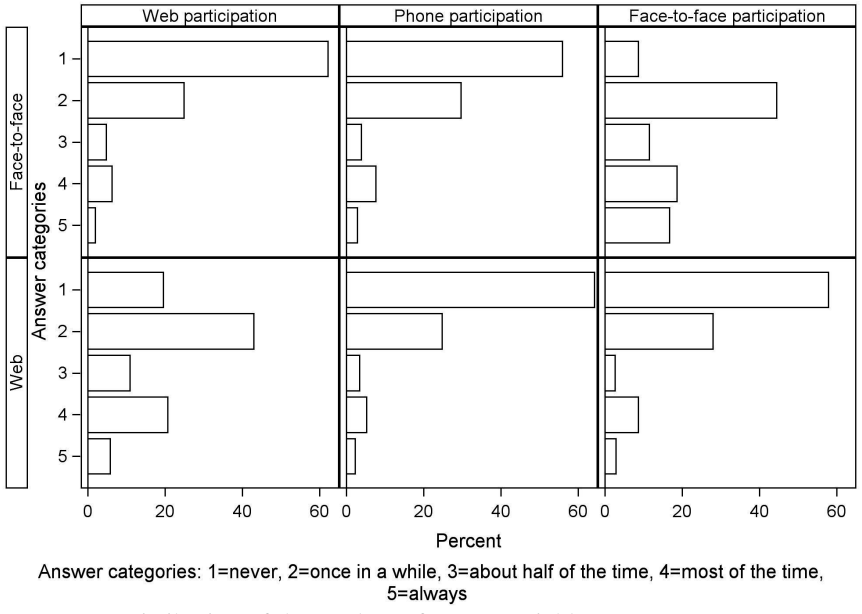

Figure 1 Distibution of the mode preference variables

We also need a set of substantive survey variables (Y) that could suffer from mode effects. We first consider measurement and selection effects on four items about survey attitudes. Although these variables were not part of the standard ESS questionnaire, but were added in the mixed mode version of the ESS in round 6, we examine these items as we expect them to suffer from strong measurement and selection effects between the web and the face-to-face mode. Indeed, these items are known to be subject to social desirability effects (negative measurement effects) (Vannieuwenhuyze et al., 2013). Moreover, the web respondents are also believed to have a more positive attitude toward surveys (positive selection effects) because they were 'early' respondents who did not require the face-to-face follow-up to

Table 2 Means and standard deviations of the variables about attitudes toward surveys for web respondents and for face-to-face respondents

Variable	Web mean	Web standard deviation	Face-to face mean	Face-to-face standard deviation
Privacy	5.11	3.10	6.51	2.96
Trust	5.19	2.43	6.35	2.64
Interest	4.39	3.08	6.28	2.75
Usefulness	6.20	2.60	6.91	2.48

participate. Therefore, we consider these 'attitude toward surveys' variables as test variables.

- Privacy (PRVCY): Do you find that surveys are an invasion of people's privacy? with the answer categories from 0 = A complete invasion of private life, to 10 = No invasion of private life at all (inverted compared with the original).
- Trust (TRSTSVY): Do you trust results obtained from a survey like this? with answer categories from 0 = No trust at all, to 10 = Complete trust.
- Interest (INTSVY): Do you find surveys like this interesting? with answer categories from 0 = Not interesting at all, to 10 = Completely interesting.
- Usefulness (USFLSVY): Do you find surveys like this useful? with answer categories from 0 = Not useful at all, to 10 = Completely useful.

Table 2 shows the means and standard deviations of these variables for the web and the face-to-face respondent groups. As expected from the social desirability hypothesis, the face-to-face respondent's means are higher than those for the web respondents.

We then consider three, four-point scale items related to attitudes toward immigration. The hypothesis for these variables is that web respondents have more positive attitudes (positive selection effects). Indeed, web respondents are in general higher educated, which is usually associated with a more positive attitude toward immigration. Moreover, the web respondents are expected to give more positive answers (positive measurement effects) due to a primacy effect caused by the vertical display of the answers in the web questionnaire, the answer category 'allow some' being read before 'allow few'. These variables are:

- Same ethnicity (IMSMETN): To what extent do you think Estonia should allow people of the same race or ethnic group as most Estonian people to come and live here?
- Different ethnicity (IMDFETN): How about people of a different race or ethnic group from most Estonian people?

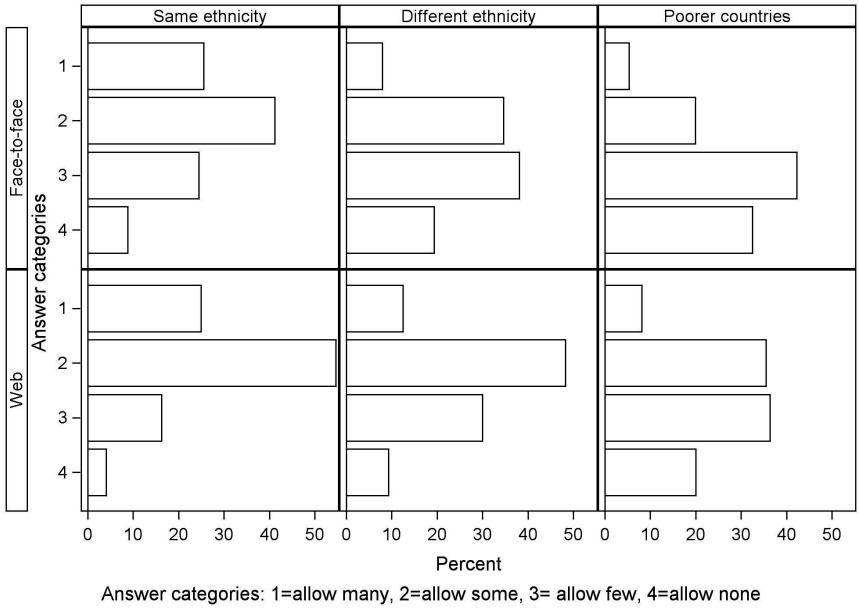

Figure 2 Distribution of the frequency of the chosen category for web respondents and for face-to-face respondents

- Poorer country (IMPRCNTR): How about people from the poorer countries outside Europe?

The answer categories are: 1= allow none, 2 = allow a few, 3 = allow some, 4 = allow many (inverted compared with the original scale).

Figure 2 shows the frequency of each answer category for these variables among web respondents and among face-to-face respondents. In this figure, the original negative scale is displayed where 1 = allow many, 2 = allow some, 3 = allow a few, and 4 = allow none. From the figure, it is clear that the category '2 = allow some' is more frequently chosen than the category '3 = allow a few' in the web questionnaire compared with the face-to-face interview.

Lastly, another set of three variables about attitudes toward immigration that have an 11-point scale rather than a four-point scale are considered.

- Economy (IMBGECO): Would you say it is generally bad or good for Estonia's economy that people come to live here from other countries? with answer categories from 0 = Bad for the economy, to 10 = Good for the economy.
- Culture (IMUECLT): And, using this card, would you say that Estonia's cultural life is generally undermined or enriched by people coming to live here from other countries? with answer categories from 0 = Cultural life is undermined, to 10 = Cultural life is enriched.

Table 3 Means and standard deviations of the variables about attitudes toward immigration for web respondents and for face-to-face respondents

Variable	Web mean	Web standard deviation	Face-to-face mean	Face-to-face standard deviation
Immigration and economy	4.98	2.30	4.66	2.52
Immigration and culture	5.45	2.53	5.43	2.44
Immigration and country	4.70	2.13	4.48	2.32

- Country (IMWBCNT): Is Estonia made a worse or a better place to live by people coming to live here from other countries? with answer categories from 0 = A worse place to live, to 10 = A better place to live.

The hypotheses for these variables are again that the web respondents have more positive attitudes (positive selection effects) toward immigration, but that the answers of the face-to-face respondents are more inclined to suffer from social desirability (negative measurement effect).

Table 3 displays the means and standard deviations of these variables for the web and for the face-to-face respondent groups. The web respondents' means are higher than those of the face-to-face respondents, despite the expected effect of social desirability, but supporting the hypothesized positive selection effect for the web.

3 Methods

The aim behind disentangling the two types of mode effects–selection and measurement–is to correct measurement effects so that results are comparable across rounds or waves of a repeated survey or across subgroups in one round. In the studied mixed mode design, the web is considered as the principal mode. Consequently, the answers given in the face-to-face interviews (observed answers) should be corrected so that they become equivalent to the answers that would have been given in a web questionnaire (counterfactual answers). To do this, we apply the back-door method, wherein a set of auxiliary variables (X) is used to model the selection effect. In the first step, the web (mode *a)* responding group is matched with the face-to-face (mode *b*) responding group through, for example, weighting. This means that the web respondents are given a weight such that the weighted web respondent group is equivalent to the face-to-face responding group, typically if considering the distribution of the set of auxiliary variables X. In the second step, the difference in estimates between the web and face-to-face respondent is split into

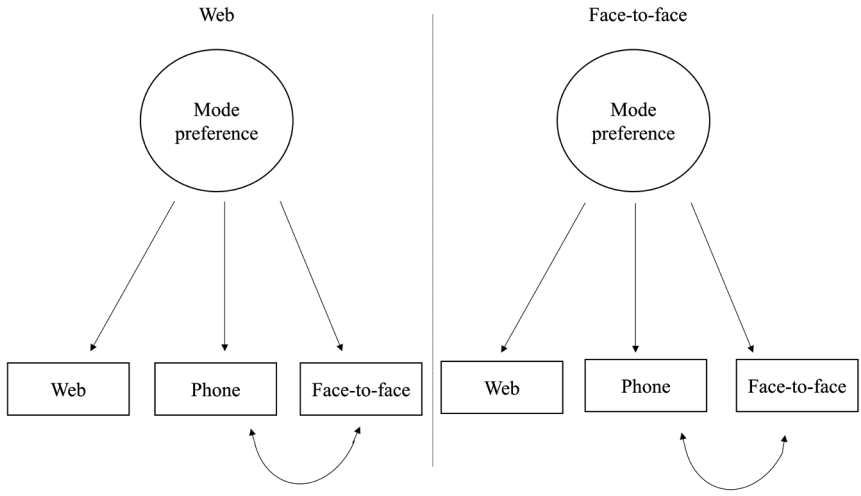

Figure 3 The mode preference latent model

(1) selection effects estimated by the difference between the web estimates and the weighted web estimates and (2) the measurement effects estimated as the difference between the weighted web estimates and the face-to-face estimates. The accuracy of the estimated measurement and selection effects depends on the compliance of the covariates (X) to the mode-insensitivity and the mode-selection ignorability assumptions.

3.1 Latent Mode Preference Variable

Given that the compliance of the set of covariates (X) with the mode-insensitivity assumption can be doubted when the mode preference variables are introduced, we create a mode preference latent variable based on the three mode-preference related variables. The construction of the latent variable should allow us to control for the measurement effect on specific items, while still extracting the essence of mode preference. A multi-group structural equation approach is applied, where the groups are defined by the modes. This approach allows us to construct equivalent latent measurement models in both modes. Because full scale equivalence between the modes appears to be too strong a requirement (CFI = 0.106, RMSEA = 0.490, and SRMR = 0.235), the equality of intercept for 'face-to-face participation' (RPF2F) and 'web participation' (RPWEB) is relaxed, but the metric equivalence and the intercept equality for the 'phone participation' (RPPHONE) are retained. The 'face-to-face participation' and 'web participation' are more likely to be subject to measurement effects between the modes than 'phone participation'. A correlation

between 'phone participation' and 'face-to-face participation' is also allowed, in order to improve the model fit (CFI = 0.993, RMSEA = 0.089, and SRMR = 0.028). This seems theoretically acceptable, as both types of data collection modes involve an interaction with an interviewer. We used the lavaan package in R to create this measurement model.

3.2 Propensity Score Weighting

We apply propensity score weighting to correct for the selection effect between the web group and the face-to-face group. The propensity score of respondent *i* is defined as the probability of *i* to participate in the web mode (mode *a*), given a set of (back-door) variables $x_1(i), x_2(i), \ldots, x_j(i)$ estimated by the logistic model (Lee & Valliant, 2008): $logit\left(p(\boldsymbol{x})\right) = \beta_0 + \beta_1 x_1 + \ldots + \beta_j x_j$.

Once estimated, in line with Lee (2006), the propensity scores are ordered and partitioned into *K* strata of equal size. We use ten strata (deciles) following the strategy shown in recent literature (Matsuo et al., 2010; Loosveldt & Sonck, 2008; Schonlau et al., 2009). If n_k denotes the total number of respondents in stratum *k*, $n_{k,b}$ the number of respondents in stratum *k* responding by a face-to-face interview (mode *b*), and $n_{k,a}$ the number of respondents in stratum *k* responding by web (mode *a*), the adjustment factor for all web respondents (mode *a*) in stratum *k* is then defined as $w_k = \frac{n_{k,b}/n_b}{n_{k,a}/n_a}$ where n_a is the total number of web respondents and n_b the total number of face-to-face respondents. This weighting scheme equates the (weighted) proportion of web respondents in stratum k with the proportion of face-to-face respondent in stratum *k*. The weighted number of web respondents in stratum *k* is given by $n_{k,a} w_k = n_{k,a} \frac{n_{k,b}/n_b}{n_{k,a}/n_a} = n_{k,b} \frac{n_a}{n_b} = n_a \frac{n_{k,b}}{n_b}$ and hence, the weighted proportion of web respondents is $\frac{n_{k,b}}{n_b}$.

3.3 The Propensity Models Based on the three Sets of Variables Considered

We calculate three sets of propensity scores in order to assess the efficiency of three different sets of back-door variables *X*.

In the first step, we used sociodemographic variables to calculate propensity scores and their associated rank strata. These variables are gender, age (4 categories: 15-29, 30-44, 45-64, and 65+), education (lower-secondary or less, upper-secondary, post-secondary or tertiary, and bachelor's, master's, or doctorate), work (in paid work or not), and geographical region of residence. In the logistic regression to estimate the propensity to participate in the web component, only education and

Table 4 Number of web and face-to-face respondents in each deciles of the propensity score distribution depending on the considered set of auxiliary variables

Deciles	Web			Face-to-face		
	Socio-demo.	+ three mode pref.	+ latent mode pref.	Socio-demo.	+ three mode pref.	+ latent mode pref.
0	11	1	10	45	54	45
1	28	3	30	27	53	26
2	30	17	27	27	39	29
3	34	31	34	22	24	21
4	38	39	35	18	17	21
5	34	45	37	22	11	19
6	38	46	39	16	9	16
7	44	53	45	11	2	12
8	40	57	44	14	0	11
9	50	55	46	7	0	9

age significantly contributed to the model. Nonetheless, all variables were retained in the logistic model in line with Lee and Valliant (2008: p. 178). The analyses were repeated with only significantly contributing variables, without implications for the results or the conclusions.

In the second step, the three mode preference variables were included, from which only 'web participation' and 'face-to-face participation' significantly contributed to the model. When mode preference variables were included, the two last strata – with the highest propensities to participate to the web component – did not contain any of the face-to-face respondents. For this reason, the web respondents in these strata (n=55+57) were given a weight of 0. This is a violation of the overlap assumption of propensity score matching methodology, which states that every unit should have a non-zero probability to be attributed to any of the groups (modes). This represents a limitation of our analysis.

In the third step, in an attempt to control for possible measurement effects on the three mode-preference variables, these three variables were replaced by a mode preference latent variable in the logistic model.

Table 4 displays the number of web respondents and of face-to-face respondents in each rank stratum for the three sets of auxiliary variables. The rank strata are deciles that were created after the web and face-to-face respondents had been ordered by propensity scores.

The distributions of web and face-to-face respondents over the propensity deciles are quite similar when these deciles are based on the sets including only sociodemographic variables and sociodemographic variables together with the mode preference latent variable. By contrast, the distribution over the deciles of web and face-to-face respondents are different when these deciles are based on the set including sociodemographic variables and the three mode preference related variables. In this case, there are almost no web respondents in the first deciles and no face-to-face respondents in the last deciles.

3.4 Estimating Selection and Measurement Effects

Assuming that the variables *X* are mode-insensitive and entirely explain the selection effect, the selection effects and the measurement effects can be expressed as follows. The answer given by respondent *i* in mode *m*, which is either web, *a*, or face-to-face, *b*, to a particular item (survey attitude or attitude toward immigration) is denoted $y_{i,m}$.

Taking the sum of the web (mode *a*) respondents, the selection effect is calculated as the difference before and after weighting:

$$S_a\left(\mu(Y)\right) = \frac{\sum_{i=1}^{n_a} y_{i,a}}{n_a} - \frac{\sum_{i=1}^{n_a} w_{k,i} y_{i,a}}{n_a} = \frac{\sum_{i=1}^{n_a} y_{i,a}}{n_a} - \frac{\sum_{i=1}^{n_a} \frac{n_{k,b}/n_b}{n_{k,a}/n_a} y_{i,a}}{n_a}$$

$$= \frac{\sum_{i=1}^{n_a} y_{i,a}}{n_a} - \frac{n_a \sum_{i=1}^{n_a} \frac{n_{k,b}}{n_{k,a}} y_{i,a}}{n_a n_b} = \frac{\sum_{k=1}^{10} n_{k,a}\mu_{k,a}}{n_a} - \frac{\sum_{k=1}^{10} n_{k,b} \sum_{i=1}^{n_{k,a}} \frac{y_{i,a}}{n_{k,a}}}{n_b}$$

$$= \frac{\sum_{k=1}^{10} n_{k,a}\mu_{k,a}}{n_a} - \frac{\sum_{k=1}^{10} n_{k,b}\mu_{k,a}}{n_b},$$

where $\mu_{k,a}$ is defined as the mean of web respondents over stratum *k*.
It should be noted that these selection effects only concern whether respondents participate in the web component of the survey rather than in the face-to-face interviews. Non-respondents are not considered.

Taking the sum of the web (mode *a*) respondents and face-to-face (mode *b*) respondents, the measurement effect is calculated as the difference of weighted responses in the web respondent group measured in mode *a* (web) and the (unweighted) responses in the face-to-face respondent group measured in mode *b* (face-to-face):

$$M_a\left(\mu(Y)\right) = \frac{\sum_{i=1}^{n_a} w_{k,i} y_{i,a}}{n_a} - \frac{\sum_{i=1}^{n_b} y_{i,b}}{n_b} = \frac{\sum_{k=1}^{10} n_{k,b}\mu_{k,a}}{n_b} - \frac{\sum_{k=1}^{10} n_{k,b}\mu_{k,b}}{n_b},$$

where $\mu_{k,a} / \mu_{k,b}$ is defined as the mean of web/face-to-face respondents over stratum k.

3.5 Significance of the Selection and Measurement Effects

Because the propensity scores are based on the respondent sample and not the full population, there is a certain degree of sampling error associated with their estimation. To integrate this level of variability, we used the bootstrap method (Efron, 1979) with 500 replicates. This means that we resampled the responding sample with a replacement 500 times – so that the replicated sample is the same size as the original responding sample–and performed the full analysis, from calculating the propensity scores to estimating the measurement and selection effects for each replicate. The variance and standard error of this collection of 500 estimates of selection and measurement effects (assuming a normal distribution) are estimates of the variability of the estimated effects. The significance of the selection and measurement effects are based on these estimated standard errors.

4 Results

Our aim in this paper was to assess the performance of mode preference variables to control for selection effects, with the goal of estimating measurement effects in the face-to-face component compared with the web component in a sequential mixed mode survey.

4.1 Model Fit of the Propensity Models

Because mode preference variables are expected to better explain selection effects between the modes, propensity models including mode preference as the independent variable should be more appropriate to predict the selected mode, and should therefore lead to a better fit of the propensity model. This better fit is confirmed by the ESS data when including the three mode preference variables alongside the sociodemographic variables: The model fit strongly improves (AIC goes from 724.5 to 420.6, pseudo-R from 0.18 to 0.66). This improvement is significant according to the residual Chi-square test (Score: 140.95/69.72 with p-values <0.001 for 4 degrees of freedom for face-to-face participation and web participation respectively). These results confirm our expectation concerning the relevance of these mode preference variables. Nevertheless, including the mode-preference latent variables instead of the three raw variables does not lead to an improvement of the model fit.

The difference in model fit improvement when using the three mode-preference variables or when using the corresponding latent variable might be an indica-

tion that the strong relationship between mode preference and the mode of participation may be explained by a violation of the mode-insensitivity assumption. Because of measurement effects on the mode-preference variables themselves, the relationship between mode preference and mode of participation may be highly overestimated.

4.2 Correlation of Propensity Scores with Target Variables

Ideal weighting variables should not only correlate with the propensity to participate in the web component of the survey, but also with the target variables (Groves, 2006; Little & Vartivarian, 2003, 2005; Kalton & Flores-Cervantes, 2003; Kalton & Maligalig, 1991; Little, 1986). As our estimated propensity scores were used to construct our weighting strata, Spearman correlation coefficients were estimated between the different target variables (attitudes toward surveys and toward immigration) and the propensity scores based on different sets of covariates (Table 5).

When considering the propensity scores based on the three mode-preference variables, results yield reduced correlation between the propensity score and the target variables. Hence, even though the mode-preference variables improve the propensity model fit of the logistic propensity model, they reduce the strength of the correlation with target variables.

When considering the propensity scores based on the latent mode-preference variable, results yield similar correlations between the propensity score and the target variables compared with when considering the propensity score based only on sociodemographic variables: Slightly stronger for attitudes toward surveys and lower for attitudes toward immigration when the mode-preference latent variable is added.

Looking at the sign of the correlations, unexpected negative correlations between the propensity score and the attitudes toward surveys (privacy, trust, interest, and usefulness) should be noted. Indeed, as web respondents are in general higher educated, and furthermore, 'early' respondents in the sequential mixed mode surveys, we expect them to have more positive attitudes toward surveys. Hence, we expect a higher propensity to participate in the web survey to be positively correlated with survey attitudes, and not negatively. A possible explanation for this surprising result is measurement effects on the surveys attitude variables causing face-to-face respondents to give more positive answers due to the presence of an interviewer.

4.3 Estimation of Measurement and Selection Effects

The effect of including mode-preference items in the propensity model to detect selection and measurement effects is central to our paper. Table 6 shows the

Table 5 Spearman correlations between target variables and propensity score for the different propensity models

Variables	sociodemo.	+ three mode pref.	+ latent mode pref.
Privacy	-0.12 ***	-0.09 *	-0.18 ***
Trust	-0.09 *	-0.09 *	-0.12 ***
Interest	-0.16 ***	-0.11 **	-0.23 ***
Usefulness	-0.12 ***	-0.05	-0.19 ***
Same ethnicity	0.07 *	0.05	0.05
Different ethnicity	0.24 ***	0.22 ***	0.22 ***
Poorer countries	0.27 ***	0.15 ***	0.24 ***
Economy	0.18 ***	0.10 *	0.17 ***
Culture	0.08 *	0.07 ‘	0.07 *
Country	0.15 ***	0.12 **	0.15 ***

‘p<0.1, * p<0.05, ** p<0.01 and ***p<0.001.

unweighted means for the web respondents, the weighted means for web respondents when the three different sets of auxiliary variables are included in the propensity model, and the mean for the face-to-face respondents. The last six columns in Table 6 display the selection and measurement effects estimates using the three different sets of covariates.

The results in Table 6 partially confirm our hypothesis concerning the direction of the selection effects, which was expected to be positive for all the variables of interest. The selection effects are indeed positive, or in most cases, not significantly different from 0 (α=0.05), independent of the set of auxiliary variables. The only exception is ‘same ethnicity’ when the three mode-preference variables are included in the propensity model, which displays a negative selection effect.

Moreover, the hypothesis concerning the measurement effect on the variables of interest is also supported by the results in Table 6. The measurement effects are all negative, or not significantly different from 0, for the 11-point scale variables about attitudes toward surveys and toward immigration. Moreover the measurement effects are positive or not significantly different from 0 for the 4-point scale variables about attitudes toward immigration.

Lastly, adding the three mode-preference variables does not lead to larger positive selection effect estimates than when only considering the sociodemographic variables. By contrast, some of the positive selection effects detected with sociodemographic variables only become not significantly different from 0. Furthermore, the selection effect on ‘same ethnicity’ is estimated as negative when the three mode-preference variables are added. When the latent variable ‘mode preference’ is added to the propensity model, the estimated selection effects are similar

Table 6 Selection and measurement effects between web and face-to-face respondents when using different auxiliary variables in the propensity score model

		Weighted web mean				Sociodemo.		+ three mode pref.		+ latent mode pref.	
Variable name	Web mean	Sociodemo.	+ three mode pref.	+ latent mode pref.	Face-to-face mean	Sel.	Meas.	Sel.	Meas.	Sel.	Meas.
Privacy	5.11	5.02	6.31	5.18	6.51	0.09	-1.50*	-1.19	-0.21	-0.07	-1.33**
Trust	5.19	5.24	5.16	5.28	6.35	-0.04	-1.11*	0.03	-1.18*	-0.09	-1.06**
Interest	4.39	4.44	4.10	4.67	6.28	-0.05	-1.84*	0.29	-2.18*	-0.28	-1.61**
Usefulness	6.21	6.23	7.14	6.43	6.91	-0.02	-0.68*	-0.93	0.23	-0.22	-0.48
Same ethnicity	3.01	2.93	3.43	2.98	2.83	0.07	0.10	-0.42*	0.59*	0.03	0.15
Different ethnicity	2.64	2.46	2.43	2.49	2.31	0.18*	0.15	0.21	0.11	0.15*	0.17
Poor countries	2.32	2.07	2.35	2.13	1.98	0.24*	0.09	-0.03	0.37	0.18*	0.15
Economy	4.98	4.55	4.86	4.59	4.67	0.43*	-0.12	0.12	0.19	0.39*	-0.08
Culture	5.45	5.08	4.88	5.33	5.43	0.38	-0.36	0.57	-0.55	0.13	-0.11
Country	4.70	4.28	4.12	4.41	4.48	0.42*	-0.20	0.58	-0.36	0.29	-0.07

*$p<0.05$

to selection effects estimated when only sociodemographic variables are considered in the propensity mode. These results are not in line with our expectations that the inclusion of mode preference in the propensity model would help to detect larger positive selection effects. There is no real pattern in the influence on selection effects of introducing the three variables concerning mode preference, showing that the measurement effects on these variables interfere greatly with the estimation of the propensity scores. The 'true' selection effects of the web compared with the face-to-face component are, however, unknown. Moreover, the overlap assumption of the propensity methodology is violated here, some web respondents could not be matched to face-to-face respondents, which could have unexpected consequences. Therefore, we are limited in drawing conclusions about the performance of the mode preference as a covariate to estimate selection effects.

5 Conclusion

The aim of this research was to test whether the inclusion of mode-preference variables in a set of covariates to control for selection effects between survey modes would offer a better trade-off between compliance with the mode-selection ignorability assumption and compliance with the mode-insensitivity assumption. To draw conclusions on the usability of mode-preference variables, three set of covariates–(1) only sociodemographic variables, (2) adding three mode-related variables, and (3) adding a mode-preference latent variable–were used in a propensity score model to evaluate the participation of respondents in the web component of a mixed-mode survey. The resulting selection and measurement effects were then compared.

The main finding is that there is no evidence that including mode-preference variables in the sets of covariates leads to more accurate estimates of the selection effects. Two cases can be distinguished: (1) no pattern can be found in the consequences for the estimated selection effects of adding the three mode-preference related variables, not controlling for mode effects on these variables, and (2) estimated selection effects are not larger (in the presumed direction) when adding the latent mode-preference variable that was constructed to control for measurement effects on the mode-preference measurements. The violation of the mode-sensitivity assumptions by the mode-preference variables seems to cause an irreversible problem, leading to the non-usability of these variables as covariates in the backdoor method. Moreover, the attempt to cancel the mode-sensitivity of the mode-preference variables by the construction of a latent variable wiped out the impact of the mode-preference variables on the selection effects.

We should mention some limitations of this research. First, empirical evidence of the absence of the added value of mode preference as a covariate is limited by the relatively small sample size and by the particularities of the survey exam-

ined: Restricted to Estonia, comparing only two modes offered sequentially, and not appointed randomly. A second limitation is that the 'true' selection effects are unknown. A third limitation is the violation of the overlap assumption of propensity matching methodology when the three mode variables are added, which could affect our conclusions. More research, in an experimental context, may be necessary to generalize our findings.

Furthermore, this research highlights the presence of measurement effects between modes in different aspects. Although almost no significant measurement effect was found on the means of the variables reflecting attitudes toward immigration, large measurement effects were found on the variables reflecting attitudes toward surveys. Therefore, attitudes toward surveys are clearly the most sensitive to social desirability. Even if adding the mode-preference variables separately reduced some of the estimated measurement effects, taking the latent variables into consideration increased them again. To conclude, even if the measurement effects between the modes are probably overestimated, the present study supports their presence.

These findings point to the risk of comparing results between data collection modes. A lot remains unexplained about the answering processes of respondents in different modes and their effects on measurement error. A possible solution would be the unimode design, in which items are designed to be robust across modes (Dillman, 2000: chapter 6).

Finally, more research might be needed in order to find adequate covariates to control for selection and measurement effects between survey modes, and to study differences in response styles between modes depending on question designs.

References

Ainsaar, M., Lilleoja, L., Lumiste, K., & Roots, A. (2013). *ESS Mixed Mode Experiment Results in Estonia (CAWI and CAPI Mode Sequential Design).* Tartu: University of Tartu, ISBN 978-9985-4-0757-8.

Bimber, B. (2003). Measuring the Gender Gap on the Internet. *Social Science Quarterly*, 81(3), 1-11.

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

Dillman, D. A. (2000). *Mail and Internet surveys: The tailored design method.* New York: John Wiley.

Dillman, D. A., Smyth, J.D., & Christian, L.M. (2009). *Internet, mail and mixed-mode surveys: the tailored design method* (3rd Ed). Hoboken: Wiley.

Dillman, D. A., Phelps, G., Tortora, R., Swift, K., Kohrell, J., Berck, J., & Messer, J. (2009). Response Rate and Measurement Differences in Mixed-Mode Surveys using Mail, Telephone, Interactive Voice Response (IVR) and the Internet. *Social Science Research,* 38, 1-18.

Efron, B. (1979). Bootstrap Method: Another Look at the Jackknife. *The Annals of Statistics,* 7(1), 1-26.

Eva, G., Loosveldt, G., Lynn, P., Martin, P., Revilla M., Saris W., & Vannieuwenhuyze, J. (2010). *Assessing the Cost-Effectiveness of Different Modes for ESS Data Collection*. London: City University.

Fowler, F. J., Gallagher, P. M. , Stringfellow, V. L., Zaslavsky, A. M, Thompson, J. W., & Cleary, P. D. (2002). Using Telephone Interviews to Reduce Nonresponse Bias to Mail Surveys of Health Plan Members. *Medical Care,* 40, 190-200.

Gentry, R., & Good, C. (2008). *Offering Respondents a Choice of Survey Mode: Use Patterns of an Internet Response Option in a Mail Survey.* Paper presented at the Annual Conference of the American Association for Public Opinion Research, New Orleans. May 15-18, 2008.

Gesell, S. B., Drain, M., & Sullivan, M. P. (2007). Test of a web and Paper Employee Satisfaction Survey: Comparison of Respondents and Non-Respondents. *International Journal of Internet Science,* 2, 45-58.

Groves, R. M. (2006). Nonresponse Rates and Nonresponse Bias in Household Surveys. *Public Opinion Quarterly,* 70(5), 646-675.

Groves, R.M., & Kahn, R. (1979). *Survey by Telephone: a National Comparison with Personal Interviews.* New York: John Wiley and Sons.

Hayashi, T. (2007). The Possibility of Mixed-mode Surveys in Sociological Studies. *International Journal of Japanese Sociology,* 16, 51-63.

Heerwegh, D., & Loosveldt, G. (2011). Assessing Mode Effects in a National Crime Victimization Survey Using Structural Equation Models: Social Desirability Bias and Acquiescence. *Journal of Official Statistics,* 27, 49-63.

Holmberg, A., Lorenc, B., & Werner, P. (2010). Contact Strategies to Improve Participation via the web in a Mixed-Mode Mail and Web Survey. *Journal of Official Statistics,* 26(3), 465-480.

Jäckle, A., Roberts, C., & Lynn, P. (2010). Assessing the Effect of Data Collection Mode on Measurement. *International Statistics Review,* 78(1), 3-20.

Kalton, G., & Flores-Cevantes, I. (2003). Weighting Methods. *Journal of Official Statistics,* 19 (1), 81-97.

Kalton, G., & Maligalig, D. (1991). A Comparison of Methods of Weighting Adjustment for Nonresponse. In *Proceedings of the 1991 Annual Research Conference* (pp. 401-428). Washington DC: U.S. Bureau of the Census.

Klausch, T., Hox, J., & Schouten, B. (2015). Selection error in single- and mixed mode surveys of the Dutch general population. *Journal of the Royal Statistical Society. Series A: Statistics in Society*, 178(4), 945-961. http://doi.org/10.1111/rssa.12102

Klausch, T., Schouten, B., & Hox, J. J. (2015). Evaluating Bias of Sequential Mixed-mode Designs against Benchmark Surveys. *Sociological Methods & Research*. http://doi.org/10.1177/0049124115585362.

Kolenikov, S., & Kennedy, C. (2014). Evaluating Three Approaches to Statistically Adjust for Mode Effects. *Journal of Survey Statistics and Methodology,* 2, 126-158.

Lee, S. (2006). Propensity Score Adjustment as a Weighting Scheme for Volunteer Panel web Surveys. *Journal of Official Statistics,* 22(2), 329-349.

Lee, S., & Valliant, R. (2008). Weighting Telephone Samples using Propensity Scores. In J. Lepkowski et al. (Eds.), *Advances in Telephone Survey Methodology* (pp. 170-183). Hoboken: John Wiley and Sons.

Little, R. J. A. (1986). Survey Nonresponse Adjustments for Estimates of Means. *International Statistical Review*, 54(2), 139-157.

Little, R. J., & Vartivarian, S. (2003). On Weighting the Rates in Non-response Weights. *Statistics in Medicine,* 22(9), 1589-1599.

Little, R. J., & Vartivarian, S. (2005). Does Weighting for Nonresponse Increase the Variance of Survey Means? *Survey Methodology,* 31, 161-168.

Loosveldt, G., & Sonck, N. (2008). An Evaluation of the Weighting Procedure for an Online Access Panel Survey. *Survey Research Method,* 2(2), 93-105.

Lugtig, P. J., Lensvelt-Mulders, G. J. L. M., Frerichs, R., & Greven, F. (2011). Estimating Nonresponse Bias and Mode Effects in a Mixed-Mode Survey. *International Journal of Market Research*, 53(5), 669-686.

Matsuo, H., Billiet, J., Loosveldt, G., Berglund, F., & Kleven, Ø. (2010). Measurement and Adjustment of Non-Response Bias based on Non-Response Surveys : the Case of Belgium and Norway in the European Social Survey Round 3. *Survey Research Methods*, 4(3), 165-178. Retrieved from http://w4.ub.uni-konstanz.de/srm/article/viewFile/3774/4332

Medway, R. L., & Fulton, J. (2012). When More Gets You Less: A Meta-Analysis of the Effect of Concurrent Web Options on Mail Survey Response Rates. *Public Opinion Quarterly*, 76(4), 733-746.

Millar, M. M., O'Neill A. C., & Dillman, D. A. (2009). *Are mode Preferences Real? Technical Report 09-003.* Pullman WA: Social & Economic Sciences Research Center. Washington State University.

Millar, M. M., & Dillman, D. A. (2011). Improving Response to web and Mixed-Mode Surveys. *Public Opinion Quarterly,* 75(2), 249-269.

Miller, T. I., Miller-Kobayashi, M., Caldwell, E., Thurston. S., & Collett, B. (2002). Citizen Surveys on the Web: General Population Survey of Community Opinion. *Social Science Computer Review,* 20, 124-136.

Morgan, S. L., & Winship, C. (2009). *Counterfactuals and causal inference: methods and principles for social research*. New York: Cambridge University Press.

Olson, K., Smyth, J. D., & Wood, H. M. (2012). Does giving People their Preferred Survey Mode actually Increase Survey Participation? *Public Opinion Quarterly,*76(4), 611-635.

Pearl, J. (2009). *Causality: Model Reasoning and Interference (2nd Ed.*). New York: Cambridge University Press.

Revilla, M. (2015). Comparison of the Quality Estimates in a Mixed-Mode and a Unimode Design: an Experiment from the European Social Survey. *Quality and Quantity,* 49, 121-1238.

Schonlau, M., van Soest, A., Kapteyn, A., & Couper, M. (2009). Selection Bias in Web Surveys and the Use of Propensity Scores. *Sociological Methods & Research*, 37(3), 291-318. http://doi.org/10.1177/0049124108327128

Schouten, B., van den Brakel, J., Buelens, B., van der Laan, J., & Klausch, T. (2013). Disentangling mode-specific selection and measurement bias in social surveys. *Social Science Research,* 42(6), 1555-1570. http://doi.org/10.1016/j.ssresearch.2013.07.005

Shi, T.-H., & Fan, X. (2007). Response Rate and Mode Preferences in Web-Mail Mixed-Mode Surveys: a Meta-Analysis. *International Journal of Internet Science*, 2(1), 59-82.

Smyth, J. D., Olson, K., & Millar, M. M. (2014). Identifying Predictors of Survey Mode Preference. *Social Science Research,* 48*,* 135-144.

Smyth, J. D., Dillman D. A., Christian, L. M., & O'Neill A. C. (2010). Using the Internet to Survey Small Towns and Communities: Limitations and Possibilities in the Early 21st Century. *American Behavioral Scientist,* 53, 1423-1448.

Tarnai, J., & Paxson, M. C. (2004). Survey Mode Preference of Business Respondents. In *Proceedings of the Survey Research Section of the American Statistical Association* (pp. 4866-4872).

US Census Bureau. (2010). Design and Methodology: American Community Survey, Washington DC: US Census Bureau. Available at: https://www.census.gov/content/dam/Census/library/publications/2010/acs/acs_design_methodology.pdf

Vannieuwenhuyze, J., Loosveldt, G., & Molenberghs, G. (2010). A Method for Evaluating Mode Effects in Mixed-Mode Surveys. *Public Opinion Quarterly,* 74(5), 1027-1045.

Vannieuwenhuyze, J., Loosveldt G., & Molenberghs, G. (2012). A Method to Evaluate Mode Effects on the Mean and Variance of a Continuous Variable in Mixed-Mode Surveys. *International Statistical Review,* 80, 306-322.

Vannieuwenhuyze, J., & Loosveldt, G. (2013). Evaluating Relative Mode Effects in Mixed-Mode Surveys: Three Methods to Disentangle Selection and Measurement Effect. *Sociological Methods and Research,* 42(1), 82-104.

Vannieuwenhuyze, J., Loosveldt, G., & Molenberghs, G. (2014). Evaluating Mode Effects in Mixed-Mode Survey Data Using Covariate Adjustment Models. *Journal of Official Statistics*, 30(1), 1-21.

Voogt, R. J. J., & Saris, W. (2005). Mixed-Mode Designs: Finding the Balance between Nonresponse Bias and Mode Effects. *Journal of Official Statistics,* 21(3), 367-387.

Weisberg, H. F. (2005). *The Total Survey Error Approach: a Guide to the New Science of Survey Research*. Chicago (III): University of Chicago.

Zillien, N., & Hargittai, E. (2009). Digital Distinction: Status-Specific Types of Internet Usage. *Social Science Quarterly*, 90(2), 274-291.

methods, data, analyses | Vol. 10(2), 2016, pp. 143-166 DOI: 10.12758/mda.2016.012

The Stability of Mode Preferences: Implications for Tailoring in Longitudinal Surveys

Tarek Al Baghal & Jennifer Kelley
University of Essex

Abstract

One suggested tailoring strategy for longitudinal surveys is giving respondents their preferred mode. Mode preference could be collected at earlier waves and used when introducing a mixed-mode design. The utility of mode preference is in question, however, due to a number of findings suggesting that preference is an artefact of mode of survey completion, and heavily affected by contextual factors. Conversely, recent findings suggest that tailoring on mode preference may lead to improved response outcomes and data quality. The current study aims to ascertain whether mode preference is a meaningful construct with utility in longitudinal surveys through analysis of data providing three important features: multiple measurements of mode preference over time; an experiment in mode preference question order; and the repeated measures within respondents collected both prior and after the introduction of mixed-mode data collection. Results show that mode preference is not a stable attitude for a large percentage of respondents, and that these responses are affected by contextual factors. However, a substantial percentage of respondents do provide stable responses over time, and may explain the positive findings elsewhere. Using mode preference to tailor longitudinal surveys should be done so with caution, but may be useful with further understanding.

Keywords: Mode preference; tailoring; mixed-mode designs; question order; context effects

1 Introduction

Obtaining survey responses across time in a longitudinal study leads to unique data collection issues compared to a cross-sectional survey, but there are also unique aspects of the longitudinal design that can be used to the benefit of the study. For example, the survey design can be adapted by taking advantage of information about respondents' life and preferences collected at earlier waves to tailor to the individual at subsequent waves. Doing so may positively influence survey outcomes through reducing burden and/or increasing interest in the survey (e.g. Lynn 2013). Examples of this tailoring include using wording that is more relevant to respondents' current situation in pre-survey mailings (Lynn 2014) or inclusion of questions of particular interest to the respondent (Oudejans 2012).

Accommodating panel members by interviewing them in their preferred mode may also increase the chances of response and data quality (Olson, Smyth and Wood 2012; Smyth, Olson, and Kasabian 2014). This form of tailoring may be of particular interest given that longitudinal surveys are increasingly incorporating mixed-mode designs as a cost consideration (e.g. Jäckle, Burton and Lynn 2015). Utilizing information on mode preference collected at earlier waves, when introducing mixed-modes for cost effectiveness and response rates, may also maximize data quality. For this usage to be effective requires that mode preference is an actual and stable attitude. However, previous findings suggest that responses to mode preference questions may be an artefact of the survey mode the preference question is asked in (e.g. Millar, O'Neill, and Dillman 2009).

The question about whether mode preferences are "real" or contextually-based is an important one, as the answer could determine the usefulness of such measures in design decisions. The limited understanding of mode preference exists largely because preference has only been asked to respondents at one point in one mode. There have not been multiple mode preference measures from the same respondent at different times, and how these measures change as the mode the respondent completes the survey may also change. This paper aims to answer questions about stability of mode preference taking advantage of a longitudinal survey providing three important features: the repeated collection of mode preference from the same individuals over time; an experiment in mode preference question order; and the repeated measures being collected both prior and after the introduction of an experiment on mixed-mode data collection. The stability (or lack thereof) of mode preference over time is most important in showing whether there is stability in atti-

Direct correspondence to
Tarek Al Baghal, Institute for Social and Economic Research, University of Essex, Wivenhoe Park, Colchester, Essex CO4 3SQ
e-mail: talbag@essex.ac.uk

tudes, while the effects of two different contextual factors will inform the potential utility of this construct in longitudinal studies. Specifically, mode of response is the contextual factor of most importance, given the argument that mode preference is a mode artefact. However, question order effects can add to the understanding of how context generally can influence mode preference.

1.1 What is mode preference?

A number of studies have defined mode preference based on revealed preferences; that is, respondents are given a choice between modes and whichever they choose is seen as the preferred mode (Diment and Garret-Jones 2007; Haan, Ongena and Aarts 2014; Shih and Fan 2007). Following Olson et al. (2012), the view taken here is this revealed preference is more appropriately seen as "mode choice" (generally among a constrained set of options) rather than "mode preference". Rather, mode preference is defined as a positive view towards being interviewed in that mode.

Several studies have directly asked respondents about mode preference as part of a survey questionnaire. No consistent preference has been identified, with most survey modes being preferred in at least one study. Findings have shown preferences for face-to-face surveys (Groves and Kahn 1979), telephone surveys (Olson et al. 2012), internet surveys (Millar et al. 2009; Tarnai and Paxson 2004), and mail surveys (Millar et al. 2009; Tarnai and Paxson 2004). Respondents also tend to report preferring the mode in which they are completing the survey at much higher rates than other modes (Groves and Kahn 1979; Millar et al. 2009; Tarnai and Paxson 2004).

1.2 Context Effects and Mode Preference

That mode preference is related to the mode of survey completion suggests responses may be affected by the survey context. When asked about subjective phenomena, respondents construct a representation based on both chronically-accessible and temporally-accessible information (Sudman, Bradburn, and Schwarz 1996). Context effects are more likely to arise when there is less reliance on chronically-accessible information (which is context-independent) than temporarily-accessible information (which is context-dependent) (Sudman et al. 1996). Temporarily-accessible information can lead to context effects depending on how it is used. The inclusion/exclusion model explaining context effects (Schwarz and Bless 1992; Bless and Schwarz 2010) suggests that temporarily-accessible information can either be assimilated or used as a contrast when assessing the representation of the target. Assimilation effects arise when the temporarily-accessible information is used in forming a representation of the target, while contrast effects occur where the information is used as a comparison standard for the target representation.

It is reasonable to assume many respondents have not given much or any deep thought about what mode they would like to be surveyed in. As such, chronically-accessible information will be limited for mode preference questions, and context effects may be expected as respondents rely on temporarily-accessible information. This reliance on temporarily-accessible information should also lead to more instability in reported preference, especially as the context changes. However, some respondents may have more chronically-accessible information regarding mode preference than others. For example, frequent internet users may have more information to draw upon regarding interacting with web designs (if not surveys specifically) than more infrequent users. In a longitudinal study, respondents who have been in the panel for a longer time will have more experience with the survey and may also have more developed preferences for their survey experience. In such cases, it may be that context effects are reduced, and with the increased chronically-available information there may be stability in mode preferences.

For those needing to rely more on temporarily-accessible information, there are possibly different sources that may provide context. Many studies have demonstrated how question order can provide context to a subjective measure through conversational norms (e.g. Schwarz, Strack and Mai 1991; Garbarski, Schaeffer, and Dykema 2015). If mode preference does not bring to mind chronically-accessible beliefs, expressed preference may also be affected by the placement of the mode preference measure in the questionnaire. If preceding questions bring to mind information pertinent to mode preferences, a context effect may occur. Specifically, if the preceding question(s) bring to mind information that "belongs" to the same category, such as questions directly related to attitudes towards specific modes included in the mode preference question, assimilation effects could be expected (Sudman et al. 1996). The information assimilated could be positive or negative, affecting the report of preferences towards or away from a particular mode.

As an example, asking first specifically about web surveys may lead frequent users of the internet to recall more positively related information to assimilate. Conversely infrequent users may have little to assimilate, or recall negative information related to the reasons of their infrequent use. However, as noted above, more frequent internet users may also not be as affected by this type of context due the availability of more chronically-accessible information. Survey experience may also affect the amount and content of accessible information. Those respondents who have more experience with the survey (e.g. participated in more waves) may have more developed attitudes towards their survey experiences. Further, those with more cognitive ability generally may be less affected by question order, possibly due the ability to make greater efforts to recall information (Narayan and Krosnick 1996).

While most research has focused on question ordering leading to context effects, it is more likely that a wide variety of information could be brought to

mind and result in context effects, including survey factors such as the mode of response (Smyth, Dillman, and Christian 2009). The respondent's survey behavior in the current mode may provide context and affect their response (Schwarz and Bohner 2001). Their survey experience in a particular mode will provide temporarily-accessible information, which may weigh heavily in response choice if little chronically-accessible information is available. Lacking prior beliefs, a positive survey experience could provide positive implications to bear on the mode preference question, increasing the chance the mode of data collection would be selected. Conversely, a negative survey experience could lead to negative implications being brought to mind, leading to some choice away from the mode of data collection. Given the voluntary nature of surveys, it can be expected that most experiences would at least not be overly negative, or else the survey could be terminated. The possibility that more survey experiences will be seen as positive, leading to positive temporarily-accessible information, would lead to results found in other studies that the mode of data collection is also the preferred mode.

Other aspects of the survey experience may also affect the information available to respondents when asked about mode preferences. For example, the presence of an interviewer may lead respondents to select an interviewer-administered mode (particularly the mode of administration) out of politeness. Particularly in longitudinal studies, where the same interviewer often returns to the same home at subsequent waves, the mode preference question may be perceived as an indicator of the respondent's attitude toward the interviewer. In such cases, the selecting the administered mode as the one preferred could be seen as the socially desirable response.

1.3 Stability of Mode Preferences

The above discussion suggests that mode preferences are largely the result of context effects, such as question order and mode of survey administration. However, it may be that mode preferences are a stable belief for some part the population, or at least some have less varying attitudes towards particular modes. This stability occurs when the available information brought to mind remains consistent in regards to the survey modes, and may be related to the context remaining the same, the amount of chronically-available information, and possibly attitude strength (Schwarz and Bohner 2001). That mode preference may be stable for at least some is suggested by two recent, related studies. Olson et al. (2012) find that when mode preference and mode offered match, cooperation increases for phone surveys and participation in both web and phone surveys. Using the same data, Smyth et al. (2014) find that responding in a preferred mode appears to reduce satisficing behaviors and improve data quality.

That a match between mode preference and the survey mode offered is related to positive outcomes suggests the measure's potential usefulness. Still, the authors

acknowledge that these findings are in contrast with those suggesting preference is a context-dependent measure. Indeed, the mode preference selected most often in this data, the telephone, was also the mode being used to conduct the initial interviews and ask the preference question. However, it is possible these results are driven by some subset of respondents that have real and stable mode preferences, while many other respondents are largely affected by the context.

Alternatively, it could be that mode preference is generally a stable attitude which affects survey behavior, and the initial data collection would be as affected by this preference as any later data collection would. If so, those preferring whatever mode is being used for interviews would respond at higher rates, the effect of which would manifest in questions on mode preference. Such an effect would in part explain the number of previous findings suggesting mode preference is an artefact of the mode of administration. There is a dearth of evidence that there is stability in mode preference or that it is largely context-driven, however, in part because of the type of data previously available. Previous studies have not explored how mode preference changes or remains stable over time in a survey within individuals, and have not explored possible contextual factors that may influence mode preference responses.

If respondents maintain the same response across time, unaffected by question order and in different modes, this would suggest mode preference is a stable attitude. Conversely, changes in responses over time, in relation to question order and/or modes would suggest that it is largely a context-dependent measurement. There may also be a mix of the two, whereby some respondents do display stable mode preferences, while others' responses are highly affected by the context. The following sections begin to provide needed evidence using repeated mode preference measures in a longitudinal mixed-mode design, taking advantage of a question-order experiment which adds further evidence to how context affects this measure.

2 Data and Methods

2.1 Sample

The Innovation Panel (IP) longitudinal survey is part of *Understanding Society: The United Kingdom Longitudinal Household Study*. The IP is a vehicle for experimentation regarding aspects of survey design in a longitudinal survey context. It uses a multi-stage probability sample of persons and households in England, Scotland, and Wales. At the fourth wave, fielded in 2011, a refreshment sample was also drawn. Waves are conducted annually, and interviews are attempted with all household members 16 years of age and older. Prior to Wave 5, all interviews were conducted by interviewers (all CAPI at Wave 4). At Wave 5, fielded in 2012, a

random two-thirds of sample households were allocated to a mixed-mode (MM) web and CAPI design, while the other third were administered the standard single-mode CAPI design. In the mixed-mode treatment, if any household member did not respond to the web survey within two weeks, an interviewer was sent to attempt a face-to-face interview with all non-responding household members. The same sample allocation was maintained at Wave 6 (in 2013). At the end of initially scheduled data collection period, contact was again attempted with some non-respondents, with the ability to complete the survey via a CATI survey (full details available at www.understandingsociety.ac.uk). However, few respondents completed CATI (n=8 in the presented data), and are not considered when examining mode of response.

The data on mode preference comes from the fourth through sixth waves of the IP. Response rates for these waves are calculated as completion rates among those responding at their initial wave of interview. At the initial wave, conducted in 2008, the response rate by original sample members was 51.7%. The Wave 4 completion rate amongst Wave 1 respondents was 54.7%; at Wave 5 there was a 45.9% completion rate among those who responded at Wave 1; and at Wave 6 a 45.9% completion rate among Wave 1 respondents. These completion rates produce a net response rate of 28.3% at Wave 4, 23.7 % at Wave 5, and 23.7% at Wave 6 (AAPOR RR3). For the refreshment sample, the Wave 4 response rate (their initial wave) was 48.8%, with completion rates among these of 82.0% at Wave 5, and 76.8% at Wave 6. These reinterview rates produce a net response rate of 40.0% at Wave 5 and 37.4% at Wave 6 (AAPOR RR3). Although attrition is significant, given the randomization of the experimental technique response propensity is not expected to interact with the experimental design and outcomes. That is, the random distribution of people with varying levels of response propensities to the experimental conditions suggest the results are not driven by differential non-response. As the goal is examining mode preference stability over time, only those respondents who answered the mode preference question at all three waves are examined (n=1477).

2.2 Mode Preference Measurement

In Waves 4 through 6, a set of five questions regarding mode preferences were asked. Two questions asked respondents to pick their most and least preferred modes among four modes (face-to-face, telephone, mail and web). Three additional questions asked about the likelihood of response (on a 0 to 10 scale, 0 = definitely would not do, 10 = definitely would do) for the specific modes of telephone, mail and web (complete question wordings available in Appendix A). A likelihood was not asked for face-to-face surveys, as the respondent was responding in a face-to-face survey at IP4. As such, it seemed apparent face-to-face was a mode in which they would complete a survey, and asking may seem redundant to the respondent.

The question asking about most preferred mode with four choices is the target question of analyses, as this is how mode preference is most frequently measured (Millar et al. 2009; Olson et al. 2012; Smyth et al. 2014; Tarnai and Paxson 2004). Further, given that there is not a specific question rating face-to-face surveys, for the reason noted above, a comparison of these questions does not allow a complete understanding of mode preference.

The order of the specific mode likelihood and most and least preferred mode questions were varied among two randomly assigned groups. One group was first asked the specific mode questions (always in the order of telephone, mail, web) and then the target question asking about most preferred mode among four choices, followed by least preferred mode. The other group was asked the target most preferred mode question first, followed by the least preferred mode, and then the three specific mode questions. Households were randomly assigned to one of these orderings at the fourth wave, and this ordering was maintained at subsequent waves. This experiment is another check on the possible context-dependent nature of mode preference questions. Question order effects are found when the order of specific and global assessments is changed (e.g. Schwarz, Strack, and Mai 1991). Again, these effects may be attributed to a greater reliance on temporally-accessible relative to chronically-accessible information (Sudman et al. 1996).

3 Results

3.1 Mode Preference Over Time

Mode preference, based on the frequently used target question asking for a preferred choice from four modes, is presented in Table 1 for the three waves this was asked. Across all three waves, the most preferred mode is a face-to-face interview, with a web survey the second most selected mode across all waves. Mail was preferred by a small percentage each wave, while very few expressed any preference for the telephone. However, there is substantial change in the numbers and percentage selecting each mode across waves. The percentage expressing a preference for face-to-face interviews decreased overall by 13.1 percentage points from the fourth to sixth wave, a relative decrease of 20.9%. Similarly, those selecting mail surveys decreased from 14.5% to 6.5% across the three waves, a 55.2% relative decrease. Conversely, there was a large increase in the number of people expressing a preference for web surveys, which coincides with the introduction to the survey of web as a mode of data collection. Nearly twenty percent more respondents selected web surveys at the sixth wave compared to the fourth wave, a 98% relative increase.
Overall, 39.5% of respondents (n=583) selected a different mode at the fifth wave than they selected at the fourth, and 26.7% (n=395) changed their response from the

Table 1 Mode Preference by Wave (in Percent)

	Wave 4 (2011)	Wave 5 (2012)	Wave 6 (2013)
Face-to-Face	63.0	51.2	49.9
Telephone	1.3	0.7	1.0
Mail	14.3	10.2	6.5
Web	20.3	35.1	40.2
No Preference	1.2	2.9	2.4

n=1477 for all waves

fifth to the sixth waves. These changes were made by 229 respondents (15.5% of the sample) who changed selections from both the fourth and fifth waves and the fifth and sixth waves, 354 (24.0%) who switched only at the fourth to fifth waves, and 166 (11.2%) switching only from the fifth to sixth waves. This totals 749 respondents (50.7% of the sample) who indicated different mode preferences across the three years at least one time.

Given these changes are within the same respondents across time suggests there is a large amount of instability in mode preference, and the possibility it is a context-dependent attitude. Regardless of the causes, such as switches in mode of survey completion (see section 3.3), the fact that dramatic changes in response distributions occur in the aggregate suggests that the attitude is not firmly held and likely more affected by temporally-accessible information, at least for a significant portion of the population. To further explore the possible existence of context effects in mode preference responses, we next turn to the mode preference question-order experiment.

3.2 The Ordering of Mode Preference Questions

The target mode preference question is a global evaluation, asking respondents to select one mode as preferred out of four options. Conversely, three questions asked about evaluations of specific modes (telephone, mail, web). The impact of the ordering of the global and specific measures is presented in Table 2. There is a clear question order effect, which is also found and replicated at subsequent waves. When the specific rating questions preceded the target question, more people selected CAPI as their preferred mode than when the target question was asked first. The reverse is true for selection of the web as the preferred mode in the target question; when this global question was asked first, more respondents chose the web as the preferred mode than when this question followed the specific questions.

Table 2 Mode by Preference by Question Ordering and Wave (in Percent)

	Wave 4 (2011)		Wave 5 (2012)		Wave 6 (2013)	
	Specific-Global	Global-Specific	Specific-Global	Global-Specific	Specific-Global	Global-Specific
Face-to-Face	67.6	57.8	54.0	48.1	51.9	47.6
Telephone	1.7	0.9	0.8	0.6	1.2	0.7
Mail	13.0	15.8	10.9	9.3	7.0	5.9
Web	16.7	24.4	32.2	38.3	38.3	42.3
No Preference	1.2	1.2	2.2	3.7	1.5	3.5
	$\chi^2_4 = 20.29$ p<0.05		$\chi^2_4 = 10.60$ p<0.05		$\chi^2_4 = 9.81$ p<0.05	

n=1477 for all waves

Of the three modes asked about specifically, web and face-to-face are the most affected in the target mode preference question. The specific question about web surveys was asked immediately previous to the target question in the specific-global order, while the design did not include a specific question about face-to-face surveys. This ordering appears to have brought more information about the web mode to mind which was assimilated when answering the target question. The results suggest that the additional temporarily-accessible information had negative implications (Sudman et al. 1996; Tourangeau, Rips, and Rasinski 2000), leading to fewer people choosing web as their preferred mode. It is unclear what these negative implications are; however, the fact this negative impact exists suggests the limited nature that chronically-accessible information has on mode preference (Sudman et al. 1996).

Respondents that may be expected to have more chronically-accessible information about survey modes also show similar patterns. Those who use the internet daily show the same significant order effect as those who use it less frequently (as does the combination of daily and several times a week internet users compared to less frequent users). The same significant order effect is also found at IP4 among both original sample members, who have more survey experience generally, and IP4 refreshment sample members, asked these questions upon their first experience. The persistence of this effect suggests the potential importance of temporally-accessible information, indicating the impermanence of mode preferences. However, there is some evidence that more educated respondents are less affected by the order experiment. At IP4, where mode is constant, differences are somewhat reduced and are only borderline significant among those with higher education (p=0.054). Higher educated respondents have been found less susceptible to other

question order experiments (e.g. Narayan and Krosnick 1996), suggesting the possibility that for some less effort is used to recall information leading to more reliance on the temporarily-accessible information.

Although ordering has an apparent impact on responses within a wave, the ordering of the questions does not appear to have an impact on the change of mode preferences across waves. The order affects responses to mode preference, but given the question order stays the same for respondents across waves, it is not particularly surprising the ordering does not affect change in responses. Cross-tabulations of change across waves and question order (not shown) found no effect between the fourth and fifth waves ($\chi^2_1 = 0.001$, p=0.982) or between the fifth and sixth waves ($\chi^2_1 = 0.180$, p=0.672). While response is affected within wave by question order, the change identified may be explained by other contextual factors, such as mode of response, which changed for some respondents across waves.

3.3 The Impact of Mode of Response on Mode Preference

The fourth wave was conducted in one mode (CAPI) and is the only available data point on mode preferences at this time. As in previous studies, the mode of completion was also selected as the preferred mode by the majority. However, an option to take the Innovation Panel survey via the web was given to some respondents at the fifth and sixth waves. The change in preferred mode among respondents to the web survey is the key to identifying the impact of mode of response on mode preference.

Table 3 presents the percentage of respondents switching their mode preference to web, another mode, or reporting the same mode preference by the mode experimental condition (CAPI-only or mixed-mode) and mode completed in at Wave 5. Those completing by web at Wave 5 changed modes from Wave 4 (where only CAPI was offered), and also changed their reported mode preference at much higher rates than anyone responding by CAPI. More web respondents changed their mode preference than repeated their response from Wave 4, whereas a large majority of both sets of CAPI respondents did not change their mode preference.

The percentage of those switching to web as their preferred mode is several times greater among web respondents than CAPI respondents in either mode condition. Additionally, the number switching to web as their preferred mode when responding by web (n=220) is greater than all CAPI respondents switching to web combined across conditions (n = 81), even though the number of all CAPI respondents combined (n=1005) greatly outnumbers the number of web respondents (n=472). Of the web respondents who changed their mode preference response (total 59.5%, n=281), 78.3% switched their preferred mode to web. Conversely, of those assigned to the CAPI-only condition, among those changing (34.0%, n=183), 30.1% switched to web; an even smaller percentage (21.8%) switched to web among those switching preference (25.5%, n=119) in the mixed-mode CAPI condition.

Table 3 Percent Switching of Mode Preference by Mode of Response, Waves 4 to 5

Mode of Response, Wave 5	Change to Web	Change to Other than Web Mode	No Change
CAPI-Only	10.2	23.8	66.1
MM, CAPI	5.6	20.0	74.5
MM, Web	46.6	12.9	40.5
Total	20.4 (n=301)	19.1 (n=282)	60.5 (n=894)

n = 1477 $\chi^2_4 = 301.59$ p<0.001

That a significantly smaller percentage of CAPI respondents in the mixed-mode condition switched to web as their stated mode preference than even those in the CAPI-only condition suggests that not only completion of the mode, but also simply offering an alternative mode may affect mode preference distributions.

The sixth wave web was again offered, and many fewer respondents switched mode across waves. Table 4, like Table 3, shows the amount of change in mode preferences from the previous to current wave, for Waves 5 and 6, but by the combination of modes completing the survey across both waves. While those in the CAPI-only condition could only complete a face-to-face interview, those in the mixed-mode condition could either complete via the same mode as in Wave 5 or the other offered mode. Few web respondents at Wave 5 switched their mode of response back to CAPI at Wave 6 (38 out of 472 Wave 5 web respondents). More respondents completed the web survey at Wave 6 after completing via CAPI at Wave 5 (128 of 466 Wave 5 mixed-mode CAPI respondents).

Again, the greatest amount of change in mode preferences occurs among those switching mode of completion from that of the previous wave. Similarly, among those moving to web from CAPI, more respondents changed their mode preference than repeated their response. Most of the change comes from these new web respondents switching their preferred mode response to match the survey mode of completion. Although a small number, those completing CAPI at Wave 6 after completing web at Wave 5 largely changed their responses to match the mode of completion as well; 16 of the 24 (66.7%) who changed mode preference did so by saying their preferred mode was now CAPI.

This changing by mode is not related to the question order; log-linear models of the three-way table (mode preference x question order x mode of response) find non-significant three-way interactions at both Wave 5 ($\chi^2_4 = 6.54$ p=0.162) and Wave 6 ($\chi^2_4 = 7.56$ p=0.1089). Given this lack of interaction, and the number of respondents shifting across modes of completion and switching mode preference,

Table 4 Percent Switching of Mode Preference by Mode of Response Across Waves 5 to 6

Mode of Response, Waves 5-6	Change to Web	Change to Other than Web Mode	No Change
CAPI-Only	7.4	19.9	72.7
MM, CAPI-CAPI	3.9	10.8	85.4
MM, Web-CAPI	10.5	52.6	36.8
MM, CAPI-Web	43.0	17.2	39.8
MM, Web-Web	13.7	8.1	78.1
Total	11.6 (n=171)	15.0 (n=220)	73.4 (n=1078)

n = 1469 $\chi^2_8 = 232.01$ p<0.001

Table 5 Mode Preference at Wave 6, by Mode of Response (in Percent)

	CAPI-Only	MM, CAPI	MM, Web
Face-to-Face	70.6	83.7	7.9
Telephone	0.9	0.5	0.5
Mail	5.7	5.9	7.7
Web	21.7	9.1	79.6
No Preference	1.7	0.8	4.3

n = 1462 $\chi^2_8 = 695.16$ p<0.001

Table 5 presents the overall distribution of mode preference by mode of completion at Wave 6. The first column shows that those who were offered only CAPI at all waves ended with a mode distribution similar to the initial measure at Wave 4. There are more respondents saying they prefer a face-to-face interview and less choosing mail as their preferred mode, but otherwise is a close approximation to the initial response distribution.

While the final outcome for CAPI-only respondents may be similar to the initial measure, it is important to note this occurs in spite of the greater amount of individual-level change for this group observed in Tables 3 and 4. A total of 34.0% of this group changed their preference at Wave 5 and 27.3% changed at Wave 6. That the end overall distribution shows substantially less change from the fourth to sixth waves suggests the possibility of random switching. When people do respond differently, a possible explanation of similarity of overall distribution is if change to/from a selection is largely random with approximately equal chance.

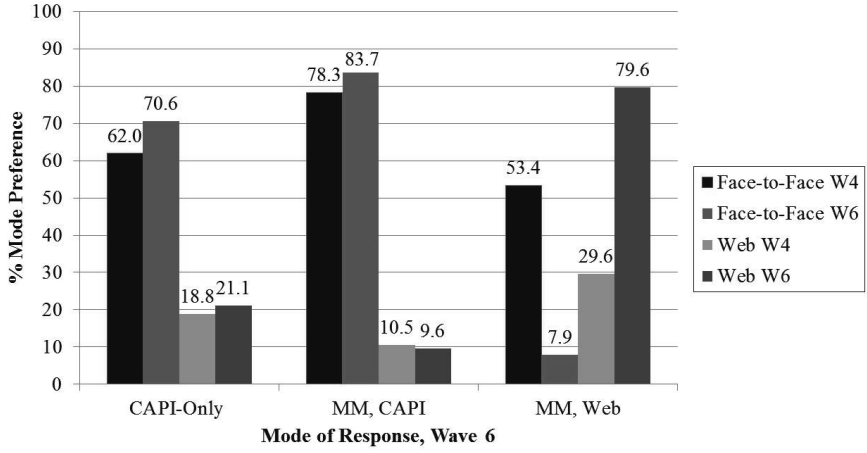

Figure 1 Mode Preference at W4 and W6 by W6 Mode of Response

Table 5 further shows that those offered web initially, but responding in CAPI, have significantly higher selection of CAPI and lower selection of web than other groups. Most striking is the shift in distribution among web respondents. Nearly 80% of web respondents selected web as their preferred mode, compared to the near 20% of the whole sample at Wave 4 or those in the CAPI-Only condition at Wave 6. Only 158 of the 444 (35.6%) of the web respondents indicating web mode preference at Wave 6 also made this choice at Wave 4. Conversely, a much smaller percentage, 7.8%, of web respondents selected a face-to-face interview as preferred; however, 53.4% of these respondents selected face-to-face as their preferred interview at Wave 4. Figure 1 displays these changes in preferences of face-to-face and web modes (the top selections) from Wave 4 and Wave 6 based on the mode of response at Wave 6.

Web respondents in particular show the drastic change in responses across waves. There was a decrease of nearly 45.5% selecting face-to-face from Wave 4 to Wave 6, while selection of web as the preferred mode increased nearly 50% in that same interval. However, there is some evidence that there is a relation between mode preference at earlier waves and mode of response at later waves. In particular, those responding in the mixed-mode CAPI version reported preferring face-to-face interviews than others at both Wave 4 and Wave 6. Similarly, those responding via the web in Wave 6 also selected a preference for web more the overall sample.

3.4 The Potential Utility and Stability of Mode Preference

The results above point to the conclusion that mode preferences are largely unstable, displaying a large amount of change across time and context. This instability and context influence suggests a possible lack of utility for these measures in designing surveys. Conversely, studies by Olson et al. (2012) and Smyth et al. (2014) stand in contrast to this conclusion suggesting the use of mode preference. A question then arises as to whether the current mode preference data, even with the large amount of change and context effects, has some relation with outcomes. Work on the use of mode preference in predicting response outcomes in the IP is ongoing, but initial results show mixed evidence that matching preference with mode offered improves response (Kaminska and Lynn 2013).

However, in a sequential mixed-mode design, as was started at the fifth wave of the IP, respondents assigned to the mixed-mode condition can respond via the web on the initial invitation. Those that do not are then offered a face-to-face alternative; in this way, respondents self-select into a mode of response (see Jackle, Burton, and Lynn 2015). Table 5 and Figure 1 provides evidence that those in the mixed-mode conditions have significantly higher levels of selecting their mode of response at Wave 6 as the preferred mode across waves modes, which may suggest stable mode preferences among some. If respondents who do prefer face-to-face surveys self-select into the CAPI-mode, while those preferring the web self-select into that mode, then the expectation would be greater percentages for the selected modes (as observed).

To explore this possibility that preference affects the selection of modes, logistic regression models were estimated predicting the mode of completion at the fifth wave of the IP, the first wave web was offered. The models include only those respondents assigned to the mixed-mode condition, as those not in the mixed-mode condition could only participate in the face-to-face survey. The models predict the probability of selecting into web response in the first model (i.e. 1= Web response, 0= Face-to-face response), and selecting into the face-to-face condition (i.e. 1= Web response, 0= Face-to-face response). The important variable and difference in the two models is the inclusion of mode preference at the previous (fourth) wave. For the web selection model, preference is indicated web versus anything else; in the CAPI completion model, preference is face-to-face versus anything else. That is, this measure is measuring in both models whether the preferred mode at the fourth wave matched the outcome variable at the fifth wave.

Since mode of response has an apparent impact on mode preference, only fourth wave responses predicting mode of response at the fifth are used, as the fourth is the last wave everyone responded in the same mode. Also included in the models are respondent characteristics of age (in years), sex (female =1), education (college or professional degree versus no higher education), race (white or

Table 6 Odds Ratios for Mode of Response at IP Fifth Wave

	Web Response	F2F Response
Daily Internet Use (at wave 4)	3.213*	0.280*
Female	1.064	0.933
College/Professional Degree	1.302	0.726
Age	0.998	1.000
Income	1.000	1.000
White	1.835*	0.573*
Employed	1.889*	0.503*
Matched Preferred Mode	2.767*	2.034*

n = 933; *p<0.001

not), employment status (employed or not) and income (measured in Great British Pounds earned per month), as well as whether the respondent used the internet daily or not. Table 6 presents the results of these models.

These basic models show that controlling for respondent demographics, previous mode preference response is strongly and positively related to which mode the respondent will select into. Respondents were more than two times more likely to select into the web survey when they stated web as the preferred mode at fourth wave, and two times more likely to select into the face-to-face survey when choosing this as their preferred mode. Internet use is also strongly related to mode selection, in an expected way. Those using the internet daily are more three times more likely to select into the web survey, while those not using the internet as frequently are estimated to be more than three and a half times (1/0.280 = 3.571) more likely to select into the CAPI survey. White and employed respondents were more likely to select into the web survey, while minorities and unemployed respondents were more likely to select into the CAPI survey.

That respondents' selection of mode is related to their previously stated mode preference suggests that the measure does predict outcomes usefully. It may be that although mode preference is affected by the context and prone to change among a large percentage of respondents, some respondents do have actual consistent mode preferences. If so, the positive results in Table 6 and found elsewhere may be driven by these consistent preferences. Indeed, while it is the case that 50.7% of respondents changed their mode preference at least once across the three waves, it also means that 49.3% of respondents gave the same mode preference response at each time point.

Understanding who has stable mode preferences could lead to better use of this measure by allowing focus on those respondents for who mode most likely matters. As initial step in understanding who is more likely to change and more likely to have stable mode preferences, Table 7 presents chi-square tests of tabula-

Table 7 Percent Reporting Same Preference Across Waves by Respondent Characteristics

	% Same Preference W4 and W5	% Same Preference W5 and W6	% Same Preference All Waves
Sex			
Females	58.4	73.0	48.4
Males	63.3	73.6	50.4
χ^2 *p-value*	0.056	0.816	0.458
Age			
<=25	50.9	66.0	34.0
25-55	60.3	73.8	48.5
55-65	57.6	72.2	46.9
>65	66.3	75.0	56.5
χ^2 *p-value*	**0.015**	0.974	**0.0004**
Education			
University/Professional Degree	60.1	73.6	48.4
Other	61.5	72.6	50.1
χ^2 *p-value*	0.596	0.664	0.344
Employment Status			
Employed	58.9	71.8	46.6
Unemployed/Not in Labor Force	61.9	74.5	51.6
χ^2 *p-value*	0.231	0.240	0.056
Income (Quartiles)			
Qt1 (lowest)	58.8	71.2	47.1
Qt2	64.1	76.6	51.4
Qt3	58.3	74.1	49.6
Qt4 (Highest)	60.8	71.1	49.1
χ^2 *p-value*	0.362	0.265	0.718
Race			
British White	60.0	73.4	48.9
Other	66.2	72.2	53.4
χ^2 *p-value*	0.163	0.769	0.322
Internet Frequency			
Every day/Several Times a week	56.3	71.1	44.7
Several times a month or less	70.2	79.2	61.5
χ^2 *p-value*	**<0.0001**	**0.002**	**<0.0001**

tion of stability in providing the same mode preference by a number of respondent characteristics (p-values less than 0.05 in bold). For example, 58.4% of females and 63.3% of males had stable mode preference from Wave 4 to 5 (Column 1). From Wave 5 to 6, mode stability increased for both females (73.0%) and males (73.6%)

(Column 2). Looking across all three waves, 48.4% of females and 50.4% reported the same mode preference each time (Column 3). However, the differences in mode preference between females and males are not significant.

Examining the other respondent demographics (age, education, employment status, income and race) by mode preference stability, from wave to wave and overall, the only significant difference in mode preferences is by age groups. Specifically, those in the oldest age group (65 years or older) are more likely to have stable mode preferences than those in the younger age categories. Internet frequency was also examined and found to have significant differences among groups; those in the less frequent internet group are more likely to have stable mode preferences than those who use the internet daily. It is not surprising that age and internet frequency both have significant differences among groups, as age and internet usages is often highly correlated. In application, it is unlikely for survey researchers to know the respondent's internet usage prior to the interview. However, age may be a viable demographic to target mode preference matches to those with more stable mode preferences.

As a further step, multivariate analyses estimating the likelihood of changing a mode preference response is estimated using multilevel logistic regression. Each respondent had two chances to change their mode preferences: between Waves 4 and 5 and between Waves 5 and 6. These models account for the dichotomous nature of the outcome variable (change or not) as well as the structure of the data as the two outcomes of change are nested within respondents. Random intercept models are used, with the one random effect occurring at the respondents-level. The outcome is set to 1 if a change in mode preference occurred across waves, 0 if the same mode preference was given. The same respondent characteristics used in Table 6 are included in this model as well. Several variables remain constant across waves sex, education (which rarely changed across waves in this data), and race. The value at the wave of interest was used for employment status, income and whether the respondent used the internet daily or not.

Two indicators for web use were tested; first was the reported internet use, the second was whether the reported internet use had changed from the previous wave. This change could indicate more or less frequent internet use, which was contrasted to those who reported the same level of internet use to the prior wave. Given that neither indicator was significant and had little impact on other findings, change in use is presented in the final models. Additionally, given the importance noted of context, a measure is included if the respondent switched mode of response across waves. Respondents could have switched to the web survey from Waves 4 to 5 and Waves 5 to 6, and could have changed from the web to face-to-face from Waves 5 to 6. To examine context further, indicators for mode of response (web or not) and the question order mode preference was asked are also included. Missing data on

Table 8 Multilevel Odds Ratio Estimates for Change in Mode Preferences

	Mode Change
Less Internet Use	0.717
More Internet Use	1.042
Income	1.000
Age	0.988*
Employed	0.779*
Female	1.154
College/Professional Degree	0.918
White	1.227
Web Mode of Response	0.741
Question Order: Asked Specific First	1.049
Switched Mode of Response	8.532*
Random-effects Parameters	
Respondent Variance	1.851
ICC	0.360

*p<0.05; Responses = 2937; Respondents = 1473

some of the independent variables lead to four respondents to be excluded from the analysis. Table 8 presents the results from this model in terms of odds ratios.

The impact of switching mode of response after controlling for all of these other factors is striking. A switch in mode of response across waves is associated with odds of switching mode preferences estimated to be eight and a half times greater than someone responding in the same mode across waves. This strong relationship in change in answers and mode supports the argument that mode preference is largely an artefact of mode of response. Other survey contextual variables are not significantly related to changes in mode preference, further indicating the importance that mode of response has on the stability of mode preference responses.

However, there is evidence that there are some respondents more likely to have stable mode preferences. In particular, those older and employed are significantly more likely to maintain a stable mode preference across waves (as indicated by lower odds of change across waves). Further, the estimated respondent intra-class correlation (ICC) suggests that respondents account for 36% of the variability in mode preference changes. This ICC shows there is still a substantial portion of variance in mode change and stability remaining relating to respondents, even after controlling for a number of survey context and respondent characteristics.

4 Discussion

The above results present evidence on the nature of mode preferences, using three aspects not explored previously: the longitudinal measurement of mode preference; the effect of changes in mode of response on mode preferences; and the impact of question order on mode preference. The results generally point in one direction – that mode preference is not stable, and is heavily influenced by contextual factors. First, examining mode preferences over time from the same respondents show significant changes at the aggregate-level across the three years it was measured. The amount of individual-level change is even greater, with more than 50% of respondents switching their response at least once in the three years the question was asked; while a substantial percentage (15%) changed responses across all three waves.

Second, the context provided by question order affects the measurement of mode preference, likely due to a lack of chronically-available information. If people do not frequently contemplate in what mode they would most like to complete a survey (which seems likely), subsequently there will be a dearth of chronically-accessible information to draw upon, increasing the opportunity for context effects (Sudman et al. 1996). That more educated respondents were less impacted by question ordering is suggestive of the availability of information theory (Narayan and Krosnick 1996). In regards to the question ordering, when the mode preference question followed the specific mode questions (immediately so by the web-specific question) more thoughts about the mode could be generated, including both positive and negative toward the attitude object (Tourangeau et al. 2000). However, face-to-face interviews were not one of the modes specifically asked about before the mode preference question. Therefore, while more positive information about the asked modes (i.e., web, telephone and mail) may already have been in active memory relative to face-to-face interviews when mode preference was asked, so would have negative information. It may be that respondents relied more on this negative information, or more negative than positive thoughts were brought to mind in the preceding questions.

Third, the mode of response also apparently provides context affecting mode preference response. Mode preference at the aggregate largely coincided with the mode of response, and changes in mode preference at the individual level are strongly related to changes in mode of response. The findings support previous assertions that mode preference is an artefact of the mode of completion (Groves and Kahn 1979; Millar et al. 2009; Olson et al. 2012; Tarnai and Paxson 2004). It may be the respondent's survey behaviors provide the contextual information to the mode preference question (Schwarz and Bohner 2001). If people do not have a definite mode preference, which the evidence presented here suggests, then survey experience will be what brings about positive or negative thoughts for the mode of

response. If the experience is neutral, there still may be a lack of negative thoughts to create a negative opinion, and selection of the mode of response as the preferred mode will at least also be consistent with their choice to respond.

These results suggest that use of mode preference to adapt and tailor longitudinal survey should be done with great caution. The lack of reliability of the measure means that decisions made on responses at one wave may be meaningless the next. There may be no gains in costs or efficiency, and could be made worse, by relying on mode preference as a tailoring strategy in mixed-mode longitudinal designs. For example, CAPI data collection at one wave appears to create a greater number of respondents selecting face-to-face interviews as their preferred mode. If at the next wave, more of the sample was allocated to CAPI, when many would have been just as likely to respond via a less expensive mode, costs could be increased over a random allocation to mode.

While the results suggest that mode preference is unstable and should be viewed with caution, studies such as Olson et al. (2012) and Smyth et al. (2014) show a number of significant positive findings suggesting the utility of allocating on mode preference. Indeed, initial analysis of the same Innovation Panel data used here shows some relation between mode preferred at Wave 4 and survey response at Wave 5 (Kaminska and Lynn 2013). It is possible that a smaller portion of respondents do have real and stable mode preferences, and these respondents are driving the positive results cited. The results presented here suggest this possibility. Mode preference is related to the selected mode when a new mode option is added at a later wave, echoing the positive results elsewhere. Further, stability in mode preference across waves is significantly related to a limited number of utilized respondent characteristics (age, employment), while a sizable portion of the remaining variance is attributable to the respondent.

The problem for longitudinal studies is identifying these respondents who actually have stable mode preference (if they exist) prior to introducing a mixed-mode design. The limited number of respondent variables identified here are not likely enough to suggest a method to identify reliable mode preferences based on individual characteristics. Further understanding of respondent characteristics related to mode preference stability is therefore needed, and may include demographics, behaviors, and other attitudes. There are also possible methods to identify mode preference reliability. For example, Cernat (2015) uses latent Markov chains to estimate reliability of measures over time and modes. The caveat to the usage of any such methods is that a longitudinal survey would have to collect several waves (at least three) of mode preference measures before these could be employed. Further research should continue to explore when and how using mode preference in longitudinal data collection is useful; however, given the observed instability in the measure, it is not clear the extensive use of mode preference will be beneficial.

References

Cernat, A. (2015). The impact of mixing modes on reliability in longitudinal studies *Sociological Methods & Research,* 44, 427-457

Bless, H., & Schwarz, N. (2010). Mental construal and the emergence of assimilation and contrast effects: The inclusion/exclusion model. *Advances in Experimental Social Psychology, 42,* 319-374.

Diment, K., and Garrett-Jones, S. (2007). How demographic characteristics affect mode preference in a postal/web mixed-mode survey of Australian researchers. *Social Science Computer Review*, 25, 410-417.

Garbarksi, D., Schaeffer N.C. & Dykema, J. (2015). The effects of response option order and question order on self-rated health. *Quality of Life Research*, 45, 1443-53

Groves, R.M. & Kahn, R.L. (1979). *Surveys by telephone*. New York: John Wiley & Sons.

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

Jäckle, A., Lynn, P. & Burton, J. (2015). Going online with a face-to-face household panel: 'effects of a mixed mode design on costs, participation rates and data quality. *Survey Research Methods,* 9, 57-70

Kaminska, O. & Lynn, P. (2013). Tailoring mode of data collection in longitudinal studies. Presented at *Hilda Survey Research Conference*, Melbourne, Australia

Lynn, P. (2013). Targeted Response Inducement Strategies on Longitudinal Surveys *Understanding Society Working Paper Series No. 2013 – 02* Institute for Social and Economic Research, University of Essex.

Lynn, P. (2014). Targeted initial letters to longitudinal survey sample members: effects on response rates, response speed, and sample composition *Understanding Society Working Paper Series No. 2014 – 08* Institute for Social and Economic Research, University of Essex.

Millar, M.M., O'Neill, A.C., & Dillman, D.A. (2009). Are mode preferences real? Technical Report 09-003 of the Social and Economic Sciences Research Center, Washington State University, Pullman, WA.

Narayan, S., & Krosnick, J. A. (1996). Education moderates some response effects in attitude measurement. *Public Opinion Quarterly*, *60*, 58-88.

Olson, K., Smyth, J.D., & Wood, H. (2012). Does giving people their preferred survey mode actually increase survey participation? An experimental examination. *Public Opinion Quarterly*, 76, 611-635.

Oudejans, M. (2012). Especially for You: Motivating Respondents in an Internet Panel by Offering Tailored Questions . Presented at the *8th International Conference on Social Science Methodology (RC33)*, Sydney, Australia.

Schwarz, N., & Bless, H. (1992). Constructing reality and its alternatives: Assimilation and contrast effects in social judgment. In Martin, L.L. & Tesser (Eds.), *The construction of social judgments* (pp. 217-245). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Schwarz, N. & Bohner, G. (2001). The Construction of Attitudes. In A. Tesser & N. Schwarz (Eds.) *Blackwell Handbook of Social Psychology: Intraindividual Processes.* Oxford, UK: Blackwell

Schwarz, N., Strack, F. & Mai, H-P. (1991). Assimilation and contrast effects in part-whole question sequences: A conversational logic analysis. *Public Opinion Quarterly,* 55, 3-23

Shih, T.-H. and Fan, X. (2007). Response rates and mode preferences in web-mail mixed-mode surveys: a meta-analysis. *International Journal of Internet Science*, 2, 59-82.

Smyth, J.D. Olson, K. and Kasabian, A. (2014). The effect of answering in a preferred versus a non-preferred survey mode on measurement. *Survey Research Methods* 8, 137-152

Smyth, J.D., Dillman, D.A., & Christian, L.M. (2009). Context effects in Internet surveys: New issues and evidence. In Joinson, A.N, McKenna, K.Y.A., Postmes, T. & Reips, U-D. (eds.) Oxford Handbook of Internet Psychology. New York: Oxford University Press.

Sudman, S., Bradburn, N.M., & Schwarz, N. (1996). *Thinking about answers.* San Francisco: Jossey-Bass.

Tarnai, J., & Paxson, M.C. (2004). Survey mode preferences of business respondents. *Proceedings of Survey Research Methods Section of the American Statistical Association*, 4898-4904.

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

Appendix A: Questions Used

Mode Preference (Target Question)

Thinking about all four ways in which we might ask you to take part in the future, including face-to-face, telephone, questionnaire sent by post or via the internet, which one would you most prefer?

1 A face-to-face interview at home

2 A telephone interview

3 A questionnaire sent by post

4 An internet questionnaire

Mode Rating Questions

Using a scale from 0 to 10 where 0 represents something you definitely would not do and 10 means something you definitely would do, if next year we approach you by telephone, how likely is it that you would complete the interview on the telephone?

And if next year we asked you to complete a paper questionnaire and return it to us by post, how likely is it that you would complete and return the questionnaire? (Presents the same scale as above).

And if next year we asked you to complete a questionnaire on the internet, how likely is it that you would complete the questionnaire? (Presents the same scale as above).

methods, data, analyses | 2016, pp. 167-194 DOI: 10.12758/mda.2016.007

Measuring the Coverage Bias in Landline Telephone Surveys by Comparison of Swiss Registry Data with Commercially Available Telephone Number Databases

Stefan Klug & Birgit Arn
DemoSCOPE

Abstract

Coverage of the population within the sampling frame is a very important quality characteristic of a study. However, a metrical evaluation of the coverage bias to approach the question of representativeness is usually not possible.

Switzerland stands out in that the federal statistical office (SFSO) has legal access to population registers (person universe) and a full list of landline telephone numbers (phone number universe). However, these data are not available for research institutes, which must rely on commercially available number collections or RDD sampling frames.

This paper wants to quantify the coverage bias of such alternative sampling frames by metric calculation of their congruence with the SFSO universes.

The analysis shows that 85.0% of private phone numbers and 88.9% of the resident population of Switzerland that can be reached via landline by the SFSO definition (non-ALTELs) are included in our exemplarily analyzed commercially available phone number collection. The highest coverage bias is present in the 20-39 age group. The RDD frame covers 97.8% of private phone numbers and 99.8% of non-ALTEL persons. Hence, both available alternative sampling frames are useful for representative studies.

Finally, the potential of use of the Swiss coverage results as benchmarks for other countries is discussed.

Keywords: CATI surveys; coverage bias; RDD; representative; Switzerland; mobile research

1 Introduction

In the early 1990s, landline penetration rates were close to saturation in most European countries (Busse et al, 2012). However, an increase in households and persons available through mobile phones only has taken place in the last decade. Additionally, willingness to publish private landline phone numbers has decreased. Due to this increase of mobile-only households and unlisted landline numbers, the usability of landline phone numbers for high quality surveys has deteriorated noticeably.

This is an international trend that is also observable in Switzerland. Studies showed that only 92% of Swiss households still had a landline phone (Stähli, 2012) and the quality of phone number samples has decreased since then; additionally, Swiss citizens are no longer obliged to list their phone numbers in the public directory. The inevitable coverage bias can lead to a significant error in the survey results, as households and persons are missing completely from the sampling frame. For example, telephone surveys imply bias related to income and household size (Stähli, 2012). According to Schouten and Bethlehem (2009), the sampling frame has to be complete to guarantee a representative response set.

In Switzerland, the SFSO uses a sampling frame called SRPH (SRPH, 2016), which contains the total resident population, for its surveys. The universe of the Swiss resident population is obtained through consolidation of municipal, cantonal and federal registers in one general data warehouse. It reflects the population at a precise reference date and is updated quarterly. SRPH therefore comes very close to a full population person and household sampling frame, as it also contains information on people living together in one household.

Additionally, the SFSO has been granted access to a list of all published and unpublished, private and business phone numbers provided by all operators in Switzerland by law[1]. This list, called Emergency Call Data Base (ECDB), repre-

1 Art. 10, Abs. 3quater of the Bundesstatistikgesetzes (SR 431.01) and Art. 16 of the Registerharmonisierungsgesetz (SR 431.02). Artikel 13a tos 13g of the Statistikerhebungsverordnung (SR 431.012.1), see also: http://www.bfs.admin.ch/bfs/portal/en/index/institutionen/oeffentliche_statistik/rechtliche_grund/bund.html

Direct correspondence to
DemoSCOPE, Klusenstrasse 17/18, 6043 Adligenswil, Switzerland
E-mail: stefan.klug@demoscope.ch / birgit.arn@demoscope.ch

Acknowledgements: We would like to thank the Swiss Federal Statistical Office (SFSO) for its great support. Our special thanks go to Claude Joye and Christoph Freymond for their excellent work and high quality analysis and contributions.
Furthermore, we thank AZ Direct and BIK Aschpurwis + Behrens for their support of our idea, and the possibility to use their data and publish the analysis results.

sents the complete universe of landline phone numbers. At the moment, mobile phone numbers are not part of the database.

SRPH can be used to sample from the complete Swiss resident population, which enables the SFSO to draw truly representative person samples. However, for a CATI survey a phone number for each survey entity is needed. Hence, in a second step, the ECDB is scanned for a phone number for each sampled person. A positive match is usually found for about 72% of the sample. If no phone number match is found, that person is called an ALTEL person by the SFSO.

However, access to SRPH and ECDB is restricted by law and not available for surveys conducted by private market and social research institutes. Data collections from commercial providers, and as a further option RDD sampling frames, must be used by market and social research agencies to conduct research by CATI surveys. Therefore, the agencies rely heavily on these other sources for sampling and the quality of these data.

Within this project, it was possible to compare ECDB and SPRH with a commercially available phone number collection and RDD samples in order to quantify the coverage bias of these sampling frames. Since the SFSO frames do not include mobile phone numbers, this analysis is restricted to landline phone numbers. Nevertheless, the range of the Swiss mobile-only penetration can be estimated from this analysis, as the maximum penetration is given by the share of persons from SRPH where no phone number match in ECDB can be found (ALTEL). As the phone number collection contains also address parameters, a match between these parameters and SRPH was also conducted in a final analysis. Note that several authors tried to access the topic of matching SRPH data and phone number collections by address parameters. However, access was always restricted to a specific sample from SRPH (Lipps et al. 2013, Lipps et al. 2015).

Coverage of a sampling frame can be defined as the percentage of landline numbers or persons in this frame that can also be found in the phone number (ECDB) of person (SRPH) universe. Coverage bias is defined as 100% coverage. As SRPH contains some socio-demographic variables, the qualitative aspects of coverage bias can be described by demographic attributes such as age or canton.

Calculation of the coverage (at a given reference date) for an alternative sampling frame allows researchers to quantify the potential lack of information and barriers to representativeness in this respect. This paper is not intended to pass judgment on ‘good’ or ‘poor’ sampling frames. Representativeness is not a dichotomous attribute: it varies from 0% to 100% and is, therefore, a quantitative measure of ‘more’ or ‘less’ representative. Ideally, a risk measure of representativeness can be calculated by multiplying the coverage of the sampling frame by the response rate of the samples. The third component – additional bias that originates from the data collection process – cannot be quantified easily and must be taken into account

as an estimate of the calculation. This simple calculation can be taken as a quantitative estimate to answer the question of the representativeness of a sample.

In the next section, we will describe examples of other sampling frames most widely used within the Swiss market and social research industry, namely data from AZ Direct (www.az-direct.ch) and a random digit dialing (RDD) frame from BIK Aschpurwis + Behrens (www.bik-gmbh.de).

The available phone number and person sources from the SFSO are described in more detail in Section 3; we will also discuss challenges associated with these SFSO sampling frames. In Section 4, we present the methodology to calculate the coverage bias. Key figures for the comparison of AZ Direct and SFSO data are shown in Sections 5 and 7. We will look at the question of whether the RDD numbers are a useful alternative in Section 6. In Section 8, we discuss the potential and conditions of use of the Swiss coverage results as a benchmark for other countries, and we attempt to analyze the added value of our results for survey researchers outside Switzerland. A general summary of the analysis is provided in Section 9.

2 Commercially Available Sources for Phone Number Samples

In the following chapter, we will describe two important sources of landline phone numbers used by survey agencies in Switzerland: a phone number collection from AZ Direct and RDD data from BIK Aschpurwis + Behrens.

2.1 AZ Direct Data

Switzerland has historically had and still has excellent landline telephone provision (Stähli, 2012). Address management companies can continuously update their databases by gathering information from a multiplicity of sources.

The most frequently used database within the market research industry is that provided by AZ Direct. This company offers a sampling frame consisting of Swiss phone numbers and a file containing data on persons and households in Switzerland. This person directory is an enriched database containing hard data and additional person and household attributes generated by means of statistical methods and data mining tools. These two sources will be labeled ‘AZ Direct Numbers’ and ‘AZ Direct Person Plus’, respectively.

Important characteristics in the AZ Direct Numbers file are the type of entry (i.e. private, business or private, and business phone number) and whether a phone number is active. This flag signals the current availability of the number in published registers. A further important feature is the language code, which allows

people or businesses to be addressed in their most likely first language. This is an important issue in a multi-language country such as Switzerland. The file also contains address information and a PersonID. Hence, it is possible to identify all numbers that belong to the same person. For all analyses in this paper, we used AZ Direct data from second quarter 2014.

The AZ Direct Numbers database consists of 5.2 million telephone numbers. Excluding numbers that are not landline numbers and keeping the Swiss numbers only (excluding Liechtenstein), we have 4.6 million landline phone numbers available for our analysis. About 3.0 million (66.0%) of these are stated as active numbers. Furthermore, private and business numbers are flagged. Note that *numbers (2.3 million, 43.8%) are not allowed to be called for marketing (i.e. sales) purposes, but can be called for market research by specific research institutes. It is a great advantage to Swiss market and social research institutes to have access to those people whose willingness to participate in CATI surveys has not been spoilt by telemarketing activities.

For 69% of entries in the AZ Direct Numbers file, AZ Direct offers additional information that can be used to restrict the selection of samples ('AZ Person Plus'). This dataset is predicated on the basis of persons rather than phone numbers. The additional information consists of address-based information, but also information on person or household attributes; e.g. the economic status of the head of the household. The PersonID is the unique link between AZ Direct Numbers and AZ Person Plus.

Note that the AZ Person Plus file cannot be used to draw a representative sample of the population. Register-based information is not accessible to private organizations such as AZ Direct and so it has to be assumed that certain selection characteristics apply to the data collection routines of AZ Direct.

2.2 RDD Sampling Frames from BIK

As an alternative to landline phone number collections, RDD (random digit dialing) offers access to a theoretically fully covered phone number sampling frame. Phone numbers in Switzerland in general are structured in such a way that the region of the landline number (or the provider of the mobile number) can be identified by the three-digit area code. Numbers can be attributed to telephony operators by number blocks and this information is publicly available. Note, however, that today a telephone number can be taken to another region or provider, and thus the system does not follow this rule any longer.

Different methods of generation of RDD numbers are described in Gabler & Häder (2007A, 2007B). Pure random digit dialing has a low hit-rate and is, therefore, inefficient. Hence, they propose a strategy in which those randomized two-digit randomization blocks are identified where at least one registered telephone

number can be found. Subsequently, every possible number in these two-digit blocks is generated. This increases the hit-rate and, furthermore, each telephone number is equally probable.

As an alternative strategy, BIK Aschpurwis + Behrens proposes that all the number blocks assigned to telephony providers in Switzerland are used. The 10,000 blocks can be downloaded from an official website (Number Blocks, 2016). Note that BIK Aschpurwis + Behrens extracted only those blocks assigned to private operators. This extraction results in a universe of 37 million phone numbers.

A further idea is to compare the performance of 10 (one-digit randomization), 100 (two-digit randomization), 1,000 (three-digit randomization) and 10,000 (four-digit randomization) randomized blocks. This means that one to four of the last digits are cut from the known blocks and complete phone numbers with all possible digit combinations are generated. This method can be applied both to the Gabler & Häder and the BIK method.

The larger the randomized block (10, 100, 1,000, 10,000), the larger the quantity of generated numbers and, hence, the larger the necessary dialing effort. However, the larger the randomization block size, the higher the coverage of the frame. So it is important to find the right trade-off between the amount of numbers and the coverage of the frame.

The dialing effort of RDD samples can be decreased through use of predictive dialing. Predictive dialing is a specific routine of the computerized dialer that predicts agent availability on interview length and other parameters. Based on this prediction, the dialer starts more calls than the number of agents that are actually available. However, predictive dialing is not a necessary prerequisite for RDD sampling. It is a potentially helpful technique that allows high quality RDD samples when sampling costs have to be reduced. For the integration of mobile phone numbers in dual-frame sampling in particular, RDD mobile sampling is the only reliable sample source, and predictive dialing is required to contact all sampled numbers within time and cost limitations (Klug et al., 2014).

See Table 1 for the comparison of the Gabler & Häder method with the BIK Aschpurwis + Behrens method. All published phone numbers in 2013 were used for the Gabler & Häder method. The number blocks for the BIK method were downloaded on July 1, 2014. The number of blocks for both methods is compared with one to four-digit randomization blocks. It can be seen that the number of blocks is always higher for the BIK method. Hence, the coverage of this method might be better than that of the Gabler & Häder method. However, a larger amount of numbers must be generated and dialed.

For the analysis in Section 6, RDD data generated by the Gabler & Häder method were used. RDD numbers have the drawback that they do not contain any address information. Although the first contact language can be roughly estimated from the area codes, true information for regional stratification is not available.

Table 1 Number of blocks for different randomization sizes, BIK and Gabler & Häder method

Block (randomization) size	# Blocks BIK method	# Blocks Gabler & Häder method
10 (one-digit)	3,728,000	820,611
100 (two-digit)	372,800	125,718
1,000 (three-digit)	37,280	19,048
10,000 (four-digit)	3,728	3,455

Based on published phone numbers, regional spreads of RDD numbers can be estimated and used for stratification.

It has been shown in numerous studies that an invitation letter together with an a-priori incentive will increase participation rates and reduce non-response bias (O'Toole et al., 2008). This is not possible with randomly generated numbers. The advantage of a potentially lower coverage bias is therefore diminished by the disadvantage of a higher non-response bias (higher as if with invitation letters).

3 SFSO Sampling Frames for Phone Numbers and Persons

The SFSO has access to a full list of landline phone numbers, including those that are not listed in public directories (ECDB). Additionally, SRPH contains a list of the Swiss resident population at a reference date.

As noted in Section 1, the list of all phone numbers is called Emergency Call Database (ECDB). Using data from the second quarter 2014, it contains about 4.1 million phone numbers, private, business and administrative. Additionally, the ECDB contains address variables and a regional/cantonal identifier for each telephone number.

The SFSO also works with a subset of the ECDB that contains all private numbers and which is relevant for population survey samples. It contains approximately 3.0 million numbers, 73.3% of all numbers in the ECDB. This SFSO sampling frame is called CASTEM (Cadre de sondage pour le tirage d'échantillons de ménages). A Venn diagram and the exact sizes of the subsets can be found in Figure 1 (left).

The identification of private numbers is made by an algorithm developed by the SFSO: all numbers where the address parameters contain a first name are judged as private numbers. This procedure might in rare cases lead to incorrect allocations.

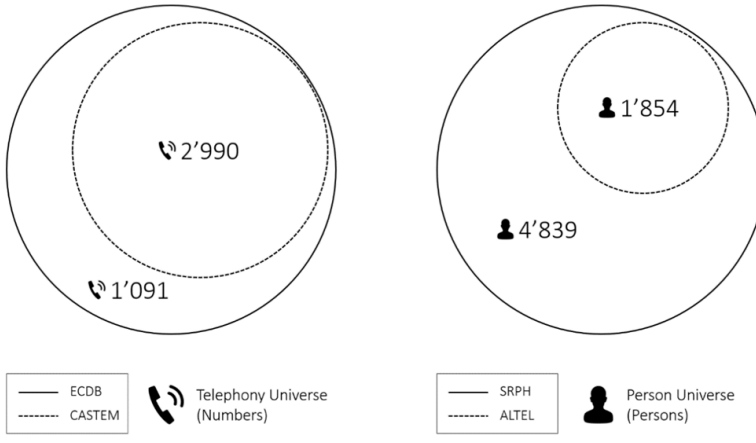

Figure 1 Telephony and person universe and the respective subsets. Numbers are in thousands and add up to the total

A final separation of business and privately used phone numbers cannot be done without direct contact with the number holder.

From 1850 to 2000, the 10-year census provided important information on the structure of the Swiss population. In 2010, a fundamental change took place. Since then, the census has been conducted and evaluated on an annual basis in a new form by the SFSO. In order to ease the burden on the population, the information is drawn primarily from population registers and supplemented by sample surveys. Only a small proportion of the population is surveyed in written form or by telephone. Thus, Switzerland now has a modern statistical system that enables observation of the development of the population and household structure, as well as the structure of buildings and dwellings.

Thanks to this new census system, the SFSO was able to build up the SRPH (SRPH, 2016). For each person in the SRPH, the following variables (in addition to name and address) are known: age, sex, language, nationality, residence permit and canton. The data from all census sources are consolidated and stored in a data warehouse (DWH), see Figure 2. This data warehouse is also the basis of the new SFSO sampling frame.

SRPH, however, does not contain phone numbers. So if a CATI survey has to be conducted, the link between SRPH and ECDB/CASTEM must be constructed. This is done by the SFSO through use of an elaborate matching algorithm that compares how many characters in the address variables of ECDB and SRPH are identical.

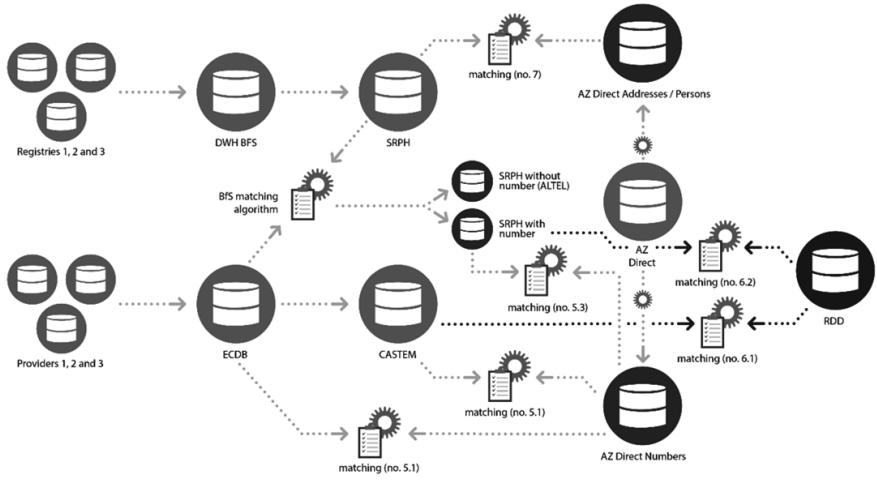

Figure 2 Schema of all data and the section (no.) with calculation of the important key figures

For data from the second quarter 2014, the number matching is possible for only 72.3% of persons in SRPH. Different reasons exist as to why a matching may not be possible (ALTEL persons); for example, if a person does not have a landline phone no match with the ECDB will be found, or the address parameters in the ECDB might be incorrect or obsolete. A Venn diagram and the exact sizes of the subsets can be found in Figure 1 (right). Variations in the matching success can be found; for example, in canton or age (see Figures 7 and 8). The matching percentage is lowest in canton Ticino (TI), whereas people living in Jura (JU) are easiest to identify in the ECDB. Furthermore, identification of phone numbers for persons aged between 20 and 39 is the most difficult. In the following sections, we distinguish between ALTEL and non-ALTEL persons. No sample from the SRPH is taken for our analysis, but we use the whole SRPH frame to compare it with a commercially available landline phone database and an RDD sample frame.

4 Comparison Methodology

Below we describe how coverage bias resulting from use of landline phone number collections or RDD samples can be calculated through comparison of these sets with available SFSO telephony and person universes.

CASTEM (as a subset to the ECDB) is the most complete collection of listed and unlisted private phone numbers. By matching a landline phone number database with CASTEM, the telephone number coverage of this phone number collec-

tion (PNC) can be calculated. In our analysis, this calculation is made using AZ Direct Numbers as an example, see Section 5.1. Telephony coverage for private numbers is then defined as $\frac{|PNC \cap CASTEM|}{|CASTEM|}$, where $|\cdot|$ is defined as the size of a set. In Section 6.1, the coverage of an RDD sample is calculated as $\frac{|RDD \cap CASTEM|}{|CASTEM|}$, using the RDD sample from BIK Aschpurwis + Behrens as an example. When calculating the telephony coverage as a key characteristic, the enumerator of the ratio is always the available telephony universe: ECDB for all phone numbers and CASTEM for all private numbers.

In this paper, SRPH defines the total population available for sampling. The coverage of other sampling sources in terms of persons can be calculated by matching them with the persons in SRPH. However, as only landline phone numbers are available, this matching can be done only for those persons in SRPH for whom a phone number can be identified. As we know from Section 3, a phone number can be found only for 72.3% of people in SRPH (non-ALTEL). Hence, for our exemplary alternative sampling sources, person coverage can be calculated as $\frac{|PNC \cap non-ALTEL|}{|SRPH|}$ and $\frac{|RDD \cap non-ALTEL|}{|SRPH|}$, respectively, see Sections 5.3 and 6.2. Note, that the size of $|PNC|$ and $|RDD|$ in this ratio is not defined as the number of phone numbers, but the number of persons. In order to obtain the total person coverage, the enumerator of these ratios must be the size of SRPH. To obtain the coverage for all persons identified by the SFSO, the enumerator can also be the size of the non-ALTELs.

ALTEL persons – as a part of SRPH – are those where no landline phone number from the ECDB can be assigned. However, this does not necessarily mean that no landline phone numbers for these persons exist. As noted above, the address parameters in ECDB connected with a phone number can be incorrect or obsolete. Hence, in Section 7 a matching is made between all SRPH persons and addresses from AZ Direct Numbers and AZ Person Plus. In this analysis, it is particularly interesting to see if the AZ Direct data can add information to the ALTELs for primary contacts via landline phone.

The calculated ratios and coverages are precise and do not need statistical correction.

Figure 2 illustrates all planned analyses and the connection between the different data sources.

5 Comparison of AZ Direct Numbers with ECDB/CASTEM and SRPH

In the following sections, we calculate the coverage of AZ Direct Numbers in terms of phone numbers (ECDB/CASTEM) and persons (SRPH).

5.1 Coverage Concerning ECDB and CASTEM

The AZ Direct Numbers collection contains 4,614,606 numbers that can be used for a match with ECDB (# numbers: 4,081,041) and CASTEM (# numbers: 2,989,632). In general, a match of sets containing numbers results in three subsets of numbers:

1. Numbers in set A only
2. Numbers in sets A and B
3. Numbers in set B only

A match of AZ Direct Numbers and ECDB shows that 1,517,319 numbers are found in AZ Direct only, which is 32.9% of all AZ Direct numbers, see Figure 3. Of these, 488,056 are flagged as active numbers, so in theory these numbers should also appear in the ECDB. A total of 3,097,287 numbers are contained in AZ Direct and ECDB frame, see Table 2 and Figure 3. Hence, 75.9% of the ECDB numbers are also represented in our exemplary landline phone number database.

If we look at the AZ Direct numbers flagged as active in detail, we see that active numbers cover 62.2% of ECDB numbers (see Table 2). This means that coverage of 13.3% (541,101 numbers) is missing, if inactive numbers are excluded from the sampling. Thus, for a market or social research company targeting high representativeness, it is important to also include numbers flagged as inactive. By extension, it is obviously good practice to provide information on formerly active numbers and keep it in the database and sampling frame. We know from previous research that people can be reached by telephone behind inactive numbers, even if at a much lower response rate than if sampling from active numbers only (Diekmann and Bruderer, 2013).

About 38.6% of the numbers flagged with an asterisk (* numbers) can be found in the ECDB. Hence, it is a clear advantage for market and social research companies that such numbers can be sampled and contacted by law.

CASTEM contains 2,989,632 numbers. Hence, the AZ Direct Numbers collection contains many more numbers than CASTEM. Note, however, that the AZ Direct database was not reduced to private numbers using the same reduction logic as for CASTEM. In CASTEM, non-private numbers are selected by filtering addresses with no first name; in the AZ Direct number database, this is done by a flag that separates business and private use of the number. Assuming a combined private and business usage of phone numbers in small businesses (which are most

Figure 3 Resulting subsets from a match of AZ Direct Numbers with ECDB and CASTEM, respectively. Numbers are in thousands and add up to the total

businesses), it makes sense to keep business numbers in the AZ Direct sampling frame and clarify usage in the interview.

The AZ Direct Numbers collection covers 85.0% of the CASTEM frame, see Table 2 and Figure 3. Note that 14.7% of the matching numbers are inactive numbers. Hence, the coverage of AZ Direct numbers is higher if we look at private phone numbers only. For the exact coverage of active and asterisk-flagged numbers and numbers with additional household information in AZ Direct, see Table 2.

Table 2 Coverage of AZ Direct Numbers in terms of ECDB, CASTEM and SRPH

		All	Flagged as active	Asterisk	With additional data
Numbers	$\lvert AZ \rvert$	4,614,606	3,044,242	2,282,966	3,607,221
	$\lvert AZ \cap ECDB \rvert$	3,097,287	2,556,186	1,575,584	2,698,647
	$\frac{\lvert \boldsymbol{AZ \cap ECDB} \rvert}{\lvert \boldsymbol{ECDB} \rvert}$	**75.9%**	**62.6%**	**38.6%**	**66.1%**
	($\lvert ECDB \rvert = 4{,}081{,}041$)				
	$\lvert AZ \cap CASTEM \rvert$	2,542,806	2,168,033	1,386,864	2,426,647
	$\frac{\lvert \boldsymbol{AZ \cap CASTEM} \rvert}{\lvert \boldsymbol{CASTEM} \rvert}$	**85.0%**	**72.5%**	**46.4%**	**81.2%**
	($\lvert CASTEM \rvert = 2{,}989{,}632$)				
Persons	$\lvert AZ \cap non-ALTEL \rvert$	4,303,048	3,810,267	2,463,010	4,161,634
	$\frac{\lvert \boldsymbol{AZ \cap non-ALTEL} \rvert}{\lvert \boldsymbol{non-ALTEL} \rvert}$	**88.9%**	**78.7%**	**50.1%**	**86.0%**
	($\lvert non-ALTEL \rvert = 4{,}838{,}986$)				
	$\lvert AZ \cap non-ALTEL \rvert$	4,303,048	3,810,267	2,463,010	4,161,634
	$\frac{\lvert \boldsymbol{AZ \cap non-ALTEL} \rvert}{\lvert \boldsymbol{SRPH} \rvert}$	**64.3%**	**56.9%**	**36.8%**	**62.2%**
	($\lvert SRPH \rvert = 6{,}693{,}298$)				

5.2 Regional Coverage

The coverage of AZ Direct numbers within ECDB and CASTEM can be further analyzed by canton (i.e. 26 regions), see Figures 4 and 5. This regional analysis is done for all numbers and the subset of active numbers. Figures 4 and Figure 5 are sorted downwards by coverage of all numbers within cantonal regions; therefore, the order varies.

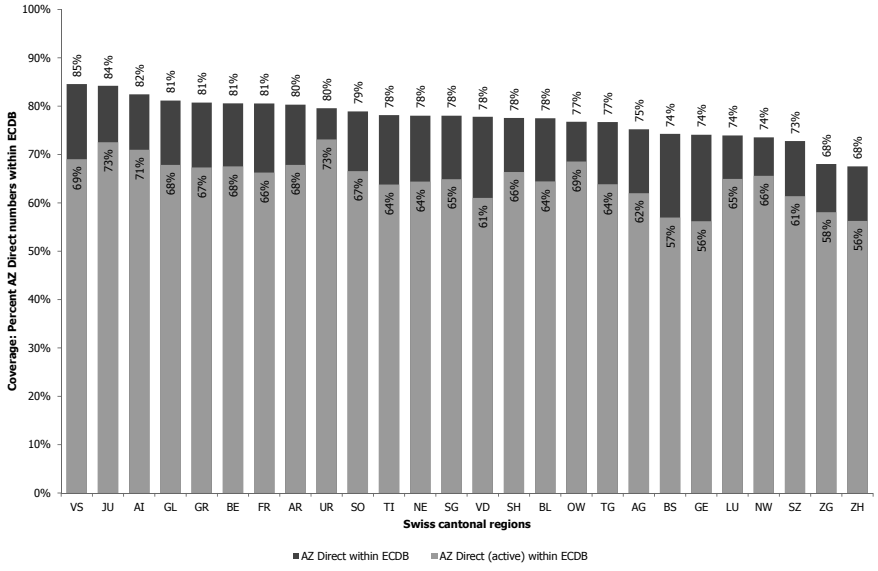

Figure 4 Coverage of AZ Direct Numbers within ECDB for all numbers and active numbers only, by canton

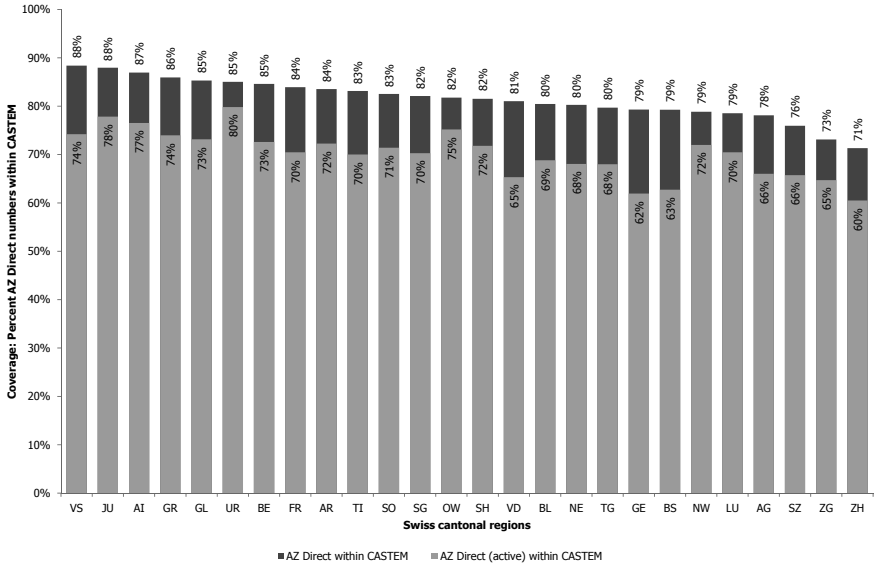

Figure 5 Coverage of AZ Direct Numbers within CASTEM for all numbers and active numbers only, by canton

It shows that coverage of AZ Direct within ECDB and CASTEM varies greatly between cantons. The coverage within CASTEM is always higher than within ECDB. Looking at all phone numbers, numbers in canton Valais (VS) and Jura (JU) are covered the most. The overall pattern for active phone numbers is similar, but on a lower level. In some cantons (Uri (UR), Nidwalden (NW), Obwalden (OW) and Lucerne (LU)), the difference between active and non-active numbers is lower (<10%) than in others.

5.3 Coverage within SRPH

When comparing AZ Direct Numbers and SRPH, the main focus is on person coverage rather than telephony coverage as before. SRPH consists of 6,693,298 persons and 3,525,438 households. As already noted in Section 3, for 27.7% of persons and 30.9% of households in SRPH, no phone number from ECDB can be matched (ALTEL), see Figure 1. This results in 4,838,986 persons (and 2,437,810 households) where a phone number can be matched (non-ALTEL).

When matching AZ Direct Numbers with SRPH persons by the assigned phone number, 88.9% of non-ALTEL persons and 88.2% of non-ALTEL households are covered. When considering the total SRPH sampling frame as the enumerator for the coverage (ALTEL and non-ALTEL), the assigned phone number of 64.3% persons (61.0% households) is part of AZ Direct Numbers. It has to be noted that this value is only approximately 8% lower than the maximum achievable value of 72.3% non-ALTEL persons. The absolute numbers are given in the Venn diagram in Figure 6.

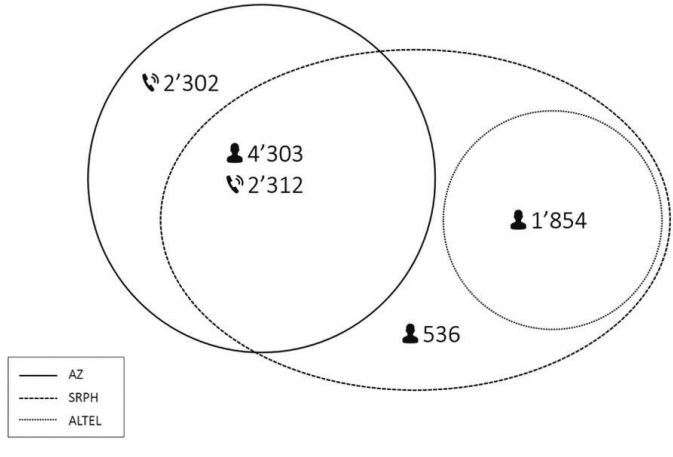

Figure 6 Resulting subsets from match of AZ Direct Numbers with SRPH. Numbers are in thousands and add up to the total

5.4 Coverage by Regional and Demographic Characteristics

Since demographic characteristics are known for the persons in SRPH, further analyses concerning coverage characteristics can be made. As the maximum achievable coverage within SRPH is the share of non-ALTEL persons, analyses show non-ALTEL coverage within SRPH in comparison with AZ Direct's key figures.

Again, the coverage within SRPH varies regionally within cantonal regions, see Figure 7. This is also reflected in coverage differences within language regions, see Figure 8. The Italian-speaking part of Switzerland has less coverage than the German and French-speaking regions. However, the low coverage can also be a result of the high share of non-permanent resident homes in this part of Switzerland and the number of Italian-speaking people who work outside their home region. In canton Ticino (TI), the SFSO faces the same challenges in identification of phone numbers within SPRH: the ALTEL share is highest in this canton (Figure 7).

Men have a lower coverage than women: 62.5% (non-ALTEL 70.1%) compared with 66.0% (non-ALTEL 72.9%). In terms of age, those aged between 20 and 39 years have the lowest AZ Direct coverage within SRPH. For higher age groups, there is almost no gap between coverage of AZ Direct Numbers and non-ALTEL persons within SRPH. The gap between non-ALTEL persons and AZ Direct Numbers is highest for those aged between 30 and 39. Thus, existing phone numbers are particularly hard for AZ Direct to collect in this age group.

For AZ Direct, foreigners living in Switzerland (AZ coverage: 45.6%, non-ALTEL: 57.6%) are harder to collect than Swiss citizens (AZ Direct coverage: 69.8%, non-ALTEL: 75.9%). The reason might be that foreigners are not as willing to publish their phone numbers in a register. The AZ Direct coverage is particularly low for holders of permit B (a time-restricted permit) (AZ coverage: 31.9%, non-ALTEL: 51%); permit C holders (permanent permit) have an AZ Direct coverage of 52% (non-ALTEL: 63%).

Not surprisingly, persons living in single-person households have the lowest coverage within the AZ Direct Numbers (see Figure 8).

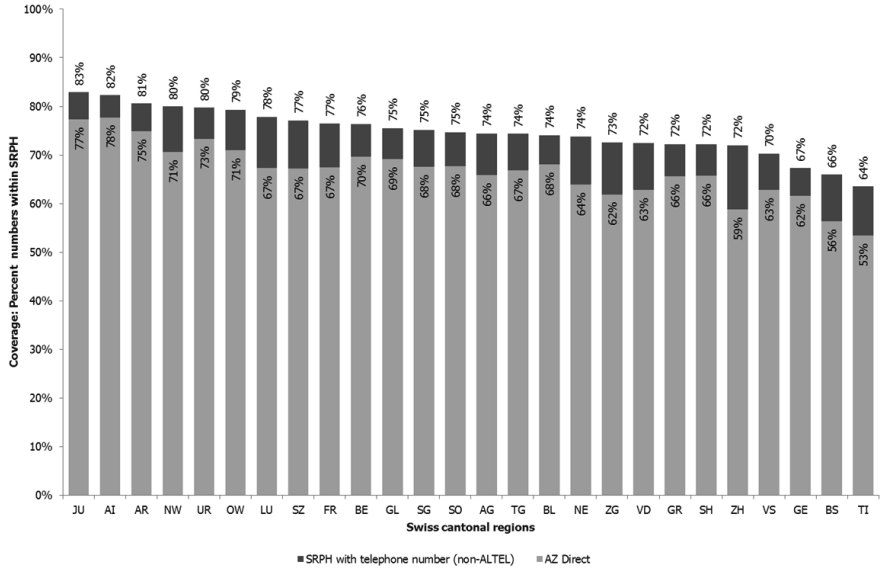

Figure 7 Coverage of AZ Direct Numbers and non-ALTEL persons within SRPH, by canton

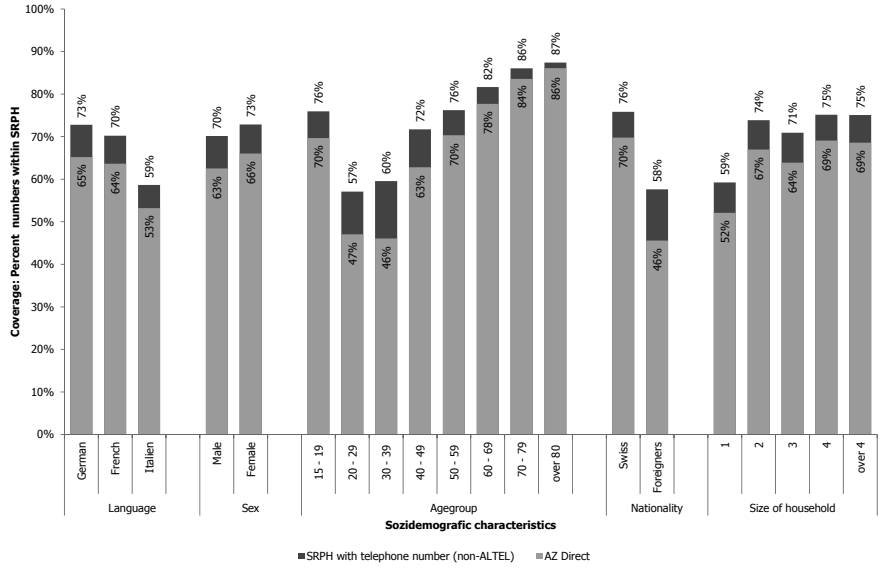

Figure 8 Coverage of AZ Direct Numbers and non-ALTEL persons within SRPH, by language region, sex, age group, nationality and size of household

6 Comparison of RDD with Different Basic Populations

The methodology used to obtain an RDD sample is described in Section 2.2. We compare RDD to CASTEM and SRPH in a similar way as AZ Direct Numbers in Sections 5.1 and 5.3, respectively. As the RDD sample in this analysis contains only private numbers, a comparison with ECDB is not discussed here.

6.1 Coverage of RDD within CASTEM

As described in Section 2.2, RDD numbers can be generated on the basis of one-digit, two-digit, three-digit or four-digit randomization. The coverage of CASTEM by the RDD sample for the different block sizes is shown in Table 3.

In theory, coverage of RDD within CASTEM has to be 100%. When creating an RDD framework, it is interesting to understand why coverage does not reach 100%. We found various explanations, all leading to the fact that valid number blocks were unknown at the time of generation of the numbers. More details are given in Section 6.2.

The coverage from two-digit randomization is 3.7% higher than for one-digit randomization. This is a significant gain in coverage, yielded by an increase of 4,365,700 phone numbers. The gain when using three-digit and four-digit randomization is not as high and many more numbers need to be generated and dialed.

The trade-off between the quantity of numbers and coverage might be best for the two-digit randomization and is also the most widely used approach in RDD sampling. For this reason, the comparison between RDD and SRPH is conducted exemplarily for the two-digit randomization in the next section.

6.2 Coverage of RDD within SRPH

In total, about 12.6 million RDD numbers are generated by the two-digit randomization, see Table 3. SRPH contains 4,838,986 persons and 2,437,810 households where a telephone number can be found in the ECDB. This was discussed in Section 5.3.

About 9.7 million telephone numbers are found in RDD only, see Figure 9. This is expected as RDD will always exceed the number of used numbers, as the approach is to capture all likely numbers by randomization. For 99.8% of non-ALTEL persons and 99.8% of non-ALTEL households, a match between the number from ECDB and the RDD sample is found. Hence, RDD provides an excellent alternative to coverage of non-ALTEL SRPH persons. In total (including ALTELs), the coverage of RDD within SRPH is 72.2% for persons and 69% for households.

Table 3 Coverage of RDD within CASTEM by different block size

Block (randomization) size	# of numbers	Coverage of RDD within CASTEM
10 (one-digit)	8,206,110	94.1%
100 (two-digit)	12,571,800	97.8%
1,000 (three-digit)	19,048,000	98.5%
10,000 (four digit)	34,550,000	98.7%

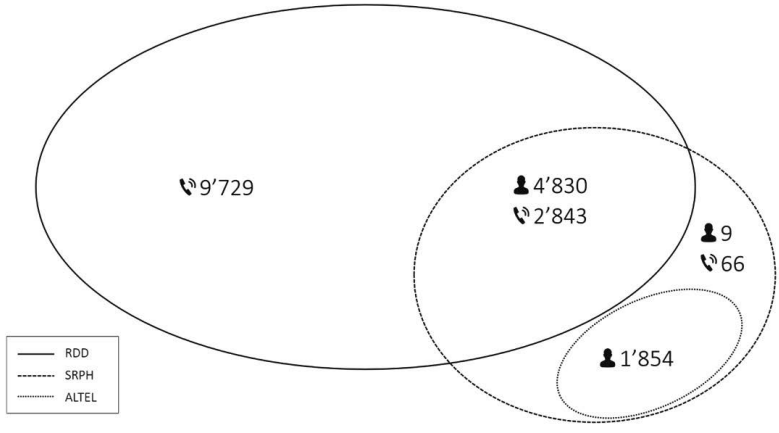

Figure 9 Resulting subsets from match of an RDD sample with SRPH. Numbers are in thousands and add up to the total

Of non-ALTEL phone numbers, no match could be found with RDD for 66,841. For 45.6% of these numbers, no entry existed in the blocks found by the Gabler & Häder method and, hence, no number was generated; i.e. this is the loss of numbers, if we use the Gabler & Häder method instead of all assigned available blocks for RDD number generation. And 54.5% of numbers were not generated because the provider was marked as a business operator (i.e. sells services to legal entities only).

Figure 10 shows the regional variability of the coverage of RDD and non-ALTEL within SRPH.

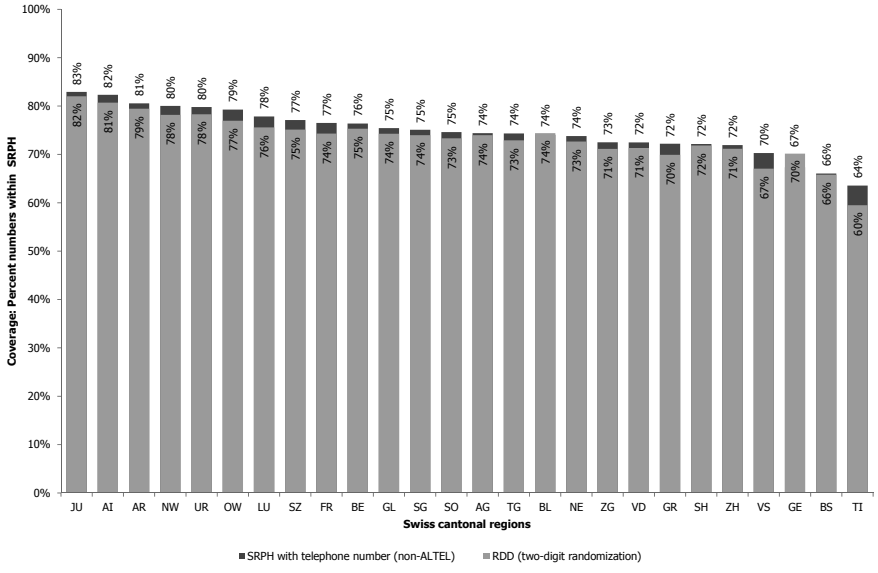

Figure 10 Coverage of RDD and non-ALTEL within SRPH, by canton

7 Supplementary Information within AZ Direct Addresses/Persons

In our last analysis, we match address items from SRPH with address items in AZ Direct to discover if the latter contains contact information by phone not in the SFSO's official phone number and person data sources. The hypothesis is that through the inclusion of not only currently activated numbers but also formerly activated (now inactive) numbers – which is unique to the AZ Direct number database – numbers can be assigned to ALTEL persons.

The question is the number of ALTEL persons for whom a landline number could be found within the AZ Direct databases using the address items (name, surname, street name, house number, postcode, place name) as matching information. In terms of an algorithm developed and tested by the SFSO, a match is defined as successful when more than 80% of the characters from the address items are identical. Usually, the SFSO conducts this matching with the data from ECDB to find numbers for SRPH persons, see Section 3. We use the same matching algorithm for the comparison between the AZ databases and SRPH.

In order to provide as much person information as possible, we use both data sources from AZ Direct, the file based on phone numbers (AZ Numbers) and the

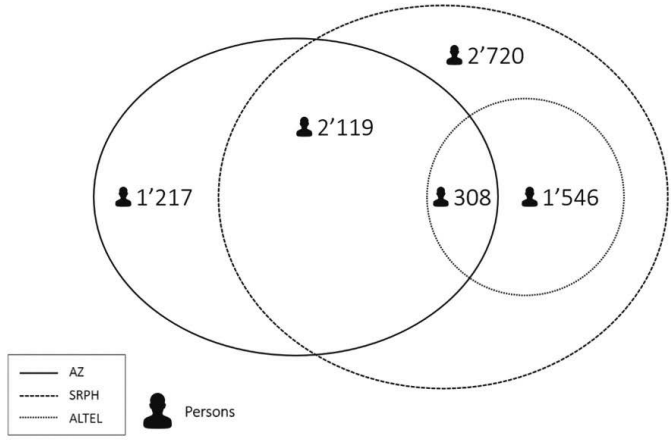

Figure 11 Resulting subsets from match of AZ Addresses/Persons with SRPH. Numbers are in thousands and add up to the total

file based on person information (AZ Person Plus). In this sense, each identifiable person within these files is linked to a number, resulting in a file structure where the telephone number is not unique. In contrast, in ECDB each number is linked uniquely to one person, the official holder of the telephone number. The resulting file is called 'AZ Addresses/Persons', see Figure 2.

In total, 3.3 million identifiable persons are within AZ Direct Addresses/Persons. SRPH contains about 6.7 million persons, see Section 5.3 and Figure 11. The large difference between these two files is not surprising, as register-based person information is, by law, not accessible to private organizations such as AZ Direct.

A match between person data from AZ Direct and SRPH is successful in 308,308 persons, but no match between SPRH and ECDB can be found (i.e. ALTEL persons), see Figure 11.

Additionally, a match can be found between person data from AZ Direct and SRPH in 50,778 addresses, but where the assigned number is different. Using such additional information is an option for the SFSO in order to generate additional landline phone contact information for CATI surveys.

A total match is found for 2,119,028 persons; i.e. 31.7% of SRPH persons. About 1,216,796 persons are in AZ Direct Addresses/Persons only and cannot be found in SRPH through use of SFSO's matching algorithm. This might be due to incorrect, incomplete or obsolete address parameters, or due to the manner in which the algorithm works. Because of the nature of this personalized information, further investigation is not possible due to data protection laws.

The analysis of the matched addresses shows only little variation in the coverage for cantons and language region. However, significant variations exist in the age group and the variable household size. One-person households are covered best by AZ Direct sources. It is not surprising that non-adult persons and non-heads of individual households cannot be found in non-official data sources. Even though AZ Direct contains some person information, it must be seen as a household or telephony register when used for population sampling.

8 Some Remarks on Potential Generalization

As noted above, Switzerland's unique legal setting facilitates the quantitative, precise calculation of coverage results, as reported in Sections 5 to 7. To our knowledge, such detailed analyses are not possible in other countries due to the lack of access to the phone number and person universe. However, a majority of the obtained results may be generalized and used as benchmarks for other countries when certain conditions apply. Depending on the fulfillment of these conditions, the calculated coverage values may not be useful as point estimates for other countries, but they could be used as upper or lower coverage levels in the country of interest.

In order to discuss the potential for generalization of the obtained results, some considerations on the datasets used in the preceding analyses must be taken into account. These are summarized in Table 4. As a general result, it can be derived that the quality and size of phone number collections (PNCs) and RDD samples depend on the percentage of listed phone numbers available to commercial providers and (for RDD samples only) the availability of published number blocks.

In order to investigate the percentage of listed phone numbers in other countries, we can look at some estimates reported by Heckel and Wiese (2012). They compared the total number of listed (published) private phone numbers with the total number of households in Germany, Italy, the UK, France and Spain, and calculated a percentage of listed phone numbers ranging from 53% to 69%. Hence, the percentage of listed phone numbers in other European countries may be lower than in Switzerland.

Sand (2014) investigated the impact of official sources of assigned number blocks for the GESIS RDD sampling frame for Germany. Depending on the availability, quality and completeness of such sources within other countries, they may or may not be used to generate RDD numbers. An overview for some European countries can be found in Heckel and Wiese (2012).

As mentioned in Table 4, census data are available for a multitude of other countries, but the quality may differ greatly. In general, the census systems can be classified as traditional, register-based, register combined with other sources, and

rolling censuses (Valente, 2010). The Swiss census belongs to the combined census type, which is also used in Italy, Germany, Spain and other central European countries. Austria and Scandinavian countries use solely register-based census systems. France is the only country to use a rolling census, whereas the UK, Portugal and most Eastern European countries use traditional census systems. To our knowledge, the use of different sampling frames within specific census types and countries is not documented and cannot be evaluated here.

Based on these preceding remarks, a discussion of the potential concerning the generalization of the results in Sections 5 to 7 in relation to other countries can be found in Table 5.

In order to assess the validity of the results concerning the unconditional person coverage (Sections 5.3 and 6.2) for other countries, the relative landline penetration must be taken into account. For a majority of European countries, this quantity is reported in Heckel and Wiese (2012, p. 111). The landline penetration in Switzerland is about 92% (Stähli, 2012).

Table 4 Remarks concerning the various datasets discussed in the preceding sections

Dataset(s)	Remarks
PNCs	▪ The conditions for commercial providers to collect (landline) phone numbers in Switzerland do not differ from conditions in other countries. The percentage of listed phone numbers may have an influence on the success of data collectors. ▪ Legal conditions and data protection laws to generate and maintain such data collections vary from country to country.
RDD samples	▪ The amount of numbers reached by using the Gabler & Häder method for number generation depends on the quality of the phone number list used as the basis for number generation (i.e. the percentage of listed phone numbers). ▪ The amount of numbers reached by using the BIK method (published number blocks) depends on the number and completeness of the published number blocks.
ECDB/ CASTEM	▪ To our knowledge, access to a complete database of published and unpublished numbers across all telephony providers for official statistics is unique to Switzerland.
SRPH	▪ Census data are available for a multitude of other countries, but the underlying census systems differ.

Table 5 Summary of discussion concerning the generalization of results in Sections 5 to 7 in relation to other countries

Section	Datasets	Remarks/discussion
5.1 & 5.2	AZ Direct Numbers vs. ECDB/CASTEM	▪ Given a similar percentage of listed phone numbers and an equivalent collecting effort/method of the commercial provider, the calculated coverage of all AZ Direct Numbers can be taken as a general benchmark for other countries. ▪ If a country has a lower percentage of listed phone numbers or the data of the commercial provider have a lower quality, the reported coverage can be seen as a maximum level. ▪ According to the results in Section 5.2, it can be taken as a general result that regional variability in coverage is present.
5.3	AZ Direct Numbers vs. SRPH	▪ In addition to the generalization conditions mentioned above, this coverage depends on the landline penetration in the country of interest. ▪ If a country has a lower landline penetration than Switzerland, the reported coverage can be seen as a maximum lower level for unconditional coverage (note that the reported coverage depends on the quality of the matching algorithm between phone numbers and registry data). ▪ The results concerning variations in coverage related to regional and demographic characteristics can be generalized at least qualitatively. For example, the finding that coverage for 30 to 39-year-olds is lowest is, in our opinion, also valid for other countries.
6.1	RDD (Häder & Gabler method) vs. CASTEM	▪ Given a similar basis for number generation as in Switzerland, the reported coverage can be taken as a general benchmark for other countries. ▪ The result that a two-digit randomization offers the best trade-off between quantity of numbers and coverage is a general result that does not depend solely on Swiss conditions.

Section	Datasets	Remarks/discussion
6.2	RDD (Häder & Gabler method) vs. SRPH	▪ In addition to the generalization conditions mentioned above, this coverage depends on the landline penetration in the country of interest. ▪ If a country has a lower landline penetration, the reported coverage can be seen as a maximum lower level for the unconditional coverage (note that the reported coverage depends on the quality of the matching algorithm between phone numbers and registry data).
7	AZ Direct Addresses/ Persons vs. SRPH	▪ As already noted in Section 2.1, we assume that certain selection criteria apply to persons in the AZ Person Plus file. Hence, we do not recommend use of the results reported in Section 7 as a benchmark for other countries or providers of data collections.

9 Conclusions

The purpose of this paper is to calculate reliable measures of coverage of alternative telephone sampling frames; i.e. commercially available alternatives to the databases available to the SFSO (ECDB/CASTEM and SRPH). The examples we use are a landline phone number collection offered by AZ Direct and RDD samples generated by BIK Aschpurwis + Behrens. The intent is not to evaluate these sources in terms of 'can be applied' or 'cannot be applied', as such a decision depends on the content, the purpose of the survey, the survey budget and other restrictions, and is finally the researchers' choice. This paper also does not include a comprehensive comparison of other methods. We assume that few options exist as far as commercial landline phone number databases are concerned. Open sources, such as internet telephone directories, cannot be used for sampling since the underlying lists or databases cannot be accessed and, therefore, randomized sample drawing is not possible.

Among the key findings here is that the exemplarily analyzed AZ Direct Numbers collection covers the population with a rate of approximately 85% concerning the telephony universe (with CASTEM as the reference population) and 64% concerning the person universe (with SRPH as the reference population). Looking at non-ALTEL persons only, the coverage within SRPH is 89%. Non-coverage is influenced by age, sex, household size and region. It must be noted that in AZ Direct Numbers, entries are missing mainly for the 20-39 age group. Additionally, the share of ALTEL within SRPH is above average within this age group. Hence,

there is a two-fold gap for this age group, which may lead to substantial bias in survey results.

When using RDD, the two-digit randomization provides the best trade-off between the quantity of generated numbers and coverage (97.8% coverage within CASTEM). In a comparison of RDD and the non-ALTEL persons, a match is found for 99.8% of persons. However, RDD has some drawbacks: an advantage of AZ Direct or SRPH over RDD samples is that households can be addressed by post before the survey starts, leading to higher response rates and, therefore, a trade-off between coverage and response rates. Also, RDD samples need predictive-dialing if research budgets are restricted.

Other non-telephony sampling approaches can be used if the risk of non-coverage bias within telephony samples appears too high for a given research target. It is clear, though, as shown in this paper, that telephone surveys still have a high measurable coverage. It can be concluded from the analyses in Sections 5 and 6 that both commercially available sources are robust sampling frames for representative studies. The choice between these two depends on researchers' risk evaluation of non-coverage against non-response and the intended study design. If no postal information is required, RDD sampling will be the preferred solution.

As discussed in Section 1, representativeness is not a decision between true or false, but studies can be representative up to a certain level. The risks and implications of a (slight) lack of representativeness can be included when the results are published.

Comparison of the AZ Direct Addresses/Persons and the SRPH addresses shows that for 308,308 ALTEL persons, a match is found within the commercially available AZ Direct database, but a further analysis of the validity and usability of this information should be considered.

Except for the analysis in Section 7, the obtained results for Switzerland can be generalized to other countries, taking into account key figures on the percentage of listed phone numbers, the availability of published number blocks (RDD only) and the landline penetration in the country of interest.

The effect on survey estimates when excluding parts of the population in telephone surveys (i.e. the coverage bias) remains an important concern among survey researchers (Massey, 1988). In future, the use of mobile numbers is essential. In particular, the two-fold gap in coverage for the 20-39 age group could be closed by the inclusion of mobile phone numbers in telephone samples. We strongly believe that as a solution dual-frame samples, including RDD mobile numbers, can bring the desired effect to high quality samples by closing coverage and overcoming the non-coverage issues shown and discussed in this paper.

Literature

Busse, B., & Fuchs, M. (2012). The Components of Landline Telephone Survey Coverage Bias. The Relative Importance of No-Phone and Mobile-Only Populations. *Quality and Quantity,* 4, 1209-1225.

Diekmann, A., & Bruderer Enzler, H. (2013). Risikosurvey 2013. Risikowahrnehmung der Schweizerinnen und Schweizer. Retrieved March 30, 2015, from ETH Zürich, Professur für Soziologie website: http://www.socio.ethz.ch/forschung/risikostudie.html

Gabler, S., & Häder, S. (2007A). Überlegungen zu einem Stichprobendesign für Telefonumfragen in Deutschland. *ZUMA-Nachrichten*, 41, 7-18.

Gabler, S., & Häder, S. (2007B). Haushalts- und Personenerhebungen – Machbarkeit von Random Digit Dialing in der Schweiz. *Methodenbericht BFS*, ISBN: 978-3-303-00378-7.

Heckel, C., & Wiese, K. (2012). Sampling Frames for Telephone Surveys in Europe. In S. Häder et al (Eds.), *Telephone Surveys in Europe* (pp. 103-119). Heidelberg: Springer.

Klug, S., Arn, B., & Müller, M. (2014). Mobile Erstkontakte in Dual-frame Stichproben aus RDD Mobile und SRPH. Schweizer Tag der öffentlichen Statistik. Bundesamt für Statistik.

Lipps, O. & Pekari, N., & Roberts, C. (2013). Coverage and nonresponse errors in an individual register frame based Swiss telephone election study. *FORS Working Papers 2013-2*.

Lipps, O. & Pekari, N., & Roberts, C. (2015). Undercoverage and Nonresponse in a List-sampled Telephone Election Survey. *Survey Research Method*s, 9 (2), 71-82.

Massey, J. T. (1988). An overview of telephone coverage. In R. M. Groves et al. (Eds.). *Telephone survey methodology* (pp. 3-8). New York: Wiley.

Number Blocks (2016). Website: http://www.eofcom.admin.ch/eofcom/public/searchEofcom_e164Allocated.do

O'Toole, J., Sinclair, M., & Leder, K. (2008). Maximising response rates in household telephone surveys. *BMC Medical Research Methodology*, 8, 71.

Sand, M. (2014). Überarbeitung des GESIS Auswahlrahmens für Telefonstichproben: Führt die Anreicherung durch die Angaben der Bundesnetzagentur zu einer Verbesserung der Auswahlgrundlage. *Dresdner Beiträge zur Soziologie*, 5, 13-38.

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

SRPH (in German: Stichprobenrahmen für Personen- und Haushaltserhebungen) (2016). Website: http://www.bfs.admin.ch/bfs/portal/de/index/news/00/08.html

Stähli, M. E. (2012). Switzerland. In S. Häder et al. (Eds.), *Telephone Surveys in Europe* (pp. 25-26). Heidelberg: Springer.

Valente, P. (2010). Census taking in Europe: how are populations counted in 2010? *Population & Societies*, 467, 1-4.

methods, data, analyses | Vol. 10(2), 2016, pp. 195-220 DOI: 10.12758/mda.2016.008

Prior Exposure to Instructional Manipulation Checks does not Attenuate Survey Context Effects Driven by Satisficing or Gricean Norms

David J. Hauser [1], *Aashna Sunderrajan* [2], *Madhuri Natarajan* [1] & *Norbert Schwarz* [3]

1 *University of Michigan*
2 *University of Illinois Urbana-Champaign*
3 *University of Southern California*

Abstract

Instructional manipulation checks (IMCs) are frequently included in unsupervised online surveys and experiments to assess whether participants pay close attention to the questions. However, IMCs are more than mere measures of attention – they also change how participants approach subsequent tasks, increasing attention and systematic reasoning. We test whether these previously documented changes in information processing moderate the emergence of response effects in surveys by presenting an IMC either before or after questions known to produce classic survey context effects. When the items precede an IMC, familiar satisficing as well as conversational effects replicate. More important, their pattern and size does not change when the items follow an IMC, in contrast to experiments with reasoning tasks. Given a power of 82% to 98% to detect an effect of $d = .3$, we conclude that prior exposure to an IMC is unlikely to increase or attenuate these types of context effects in surveys.

Keywords: instructional manipulation checks; survey context effects; satisficing; Gricean conversational norms; survey methods

1 Introduction

With the surge in cheap, fast research via online labor markets (Buhrmester, Kwang, & Gosling, 2011), the issue of participant attentiveness has received considerable attention from behavioral researchers (Goodman, Cryder, & Cheema, 2013; Berinksy, Margolis, & Sances, 2013; Paolacci, Chandler, & Ipeirotis, 2010). Some have expressed concern over participant attentiveness in online tasks (see "Quality Assurance" section in Mason & Suri, 2012). Furthermore, many researchers see it as a major issue for research conducted on online labor markets (see informal poll in Chandler, Mueller, & Paolacci, 2014).

One popular method of ensuring attention is the Instructional Manipulation Check (IMC; Oppenheimer, Meyvis, & Davidenko, 2009). The typical IMC is a question that requires close attention to the instructions in order to answer the question correctly; hence, not answering the question correctly is treated as an indication of not paying close attention to the instructions. The standard IMC on the surface looks like a humdrum survey question but contains less noticeable text in the instructions that informs participants to provide an unconventional response in place of an intuitively correct response (Oppenheimer et al., 2009). As an example, a bolded lure question might inquire about which sports you play, but hidden in the instructions may be a command to click the title of the question in order to demonstrate attention. Other methods of checking on participant attention involve asking questions with factually correct, obvious answers, such as, "While watching television, have you ever had a fatal heart attack?" Participants selecting any response other than "never" are presumed to have not been paying attention while responding (Paolacci et al., 2010). These inattentive participants often contribute substantial error to datasets by failing to read the entirety of instructions or by not giving enough thought to questions, which can justify excluding them from analyses (Oppenheimer et al., 2009). Hence, the routine use of IMCs is frequently recommended by online research methodologists as a way to validate online participant pool platforms (e.g., Paolacci et al., 2010; Goodman et al., 2013; Berinsky, Margolis, & Sances, 2013), and they have been become prevalent research tools.

Despite their prevalence as measures of attention, little research has explored how the administration of an IMC itself may affect participants' inferences about the study and their responses to a questionnaire. As research into context effects in self-report highlights, every question is also a treatment that may affect responses

Acknowledgments

We thank Allison Earl for her advice with the research.

Direct correspondence to

David J. Hauser, Department of Psychology, 3233 East Hall
530 Church St, Ann Arbor, MI 48109-1043
E-mail: djhauser@umich.edu

to subsequent questions (for reviews, see Schwarz, 1999; Sudman, Bradburn, & Schwarz, 1996). This may be particularly likely for IMCs, which stand out as unique, salient questions in the context of a standard survey. These questions usually convey the message that researchers want to know if participants are paying attention. This highlights that paying close attention and reading all instructions is important and highly valued in this survey. Furthermore, these questions often attempt to lure participants into responding incorrectly. Thus, IMCs also inform participants that questions may not be what they seem and that the survey may involve "trick" questions that should not be taken at face value. These lessons may increase attention to detail and may prompt a more systematic reasoning strategy than respondents might otherwise adopt.

Initial research suggests that this may be the case. Hauser and Schwarz (2015a, Experiment 1) had participants answer a standard IMC and complete the Cognitive Reflection Test, a series of math questions designed to measure a person's propensity to engage in reflective thinking (Frederick, 2005). For example, a question would read, "If it takes 5 machines 5 minutes to make 5 widgets, how long would it take 100 machines to make 100 widgets?" (taken from Frederick, 2005). People tend to intuitively respond "100," but the actual answer is "5," which requires more careful, reflective thinking to reach. Hauser and Schwarz varied the order of the tasks, such that the CRT questions either preceded or followed a single IMC question. As expected, participants performed better on the CRT when they had first answered an IMC question. A follow-up study further showed that answering an IMC improves performance on subsequent probabilistic reasoning tasks (Hauser & Schwarz, 2015a, Experiment 2). These findings converge on the conclusion that IMCs do more than "assess" participants' attention: they teach participants that there may be more to a question than meets the eye, which influences how they approach later questions in the survey. As a result, participants who were exposed to an IMC engage in more careful reasoning on subsequent questions, compared to participants who were not exposed to an IMC.

Whether this is a desirable or undesirable effect of using IMCs depends on the researcher's goals. If one wants the most careful answers possible, IMCs may be helpful in achieving the goal. But if one wants to capture how and what people think spontaneously, IMCs may systematically bias one's results. Using the above reasoning tasks as an example, a preceding IMC may be desirable when one wants to test how well people can do when highly motivated. Yet the sample's enhanced performance when an IMC is administered is likely to differ from the performance one would observe under many natural conditions, resulting in erroneous population estimates.

At this point, it is unknown how general the influence of IMCs is. On the one hand, IMCs may only affect performance on tasks that look "tricky" to begin with, such as complex reasoning tasks where correct responses are nonobvious and

require overriding intuitive responses. The tasks affected by IMC administration to this point have fallen into this category, so it is currently unknown whether IMCs may affect other subsequent tasks. On the other hand, participants' motivation and their assumptions about the cooperative nature of the research conversation have been shown to play a key role in all self-report tasks. For instance, minute aspects of surveys such as the survey's letterhead (Norenzayan & Schwarz, 1999), question order (Schwarz, Strack, & Mai, 1991), and administration mode (Schwarz, Strack, Hippler, & Bishop, 1991) all affect survey behavior. Thus, it seems possible that an IMC may influence many common survey tasks because of the unique information that it conveys. Next, we review survey tasks that may be particularly likely to be influenced by IMC placement, namely tasks that give rise to satisficing and Gricean conversational norm effects.

2 Satisficing

Participants often exert less than optimal effort in answering questions. Termed satisficing (Krosnick, 1991, 1999), the phenomenon refers to the practice of taking mental shortcuts rather than considering the full range of inputs in responding to survey questions. Satisficing manifests in specific patterns of survey behavior. *Response order effects* emerge when satisficing participants select the first most reasonable response, resulting in different responses when response option order is manipulated (Schuman & Presser, 1981). Satisficing participants also display *non-differentiation* (Krosnick, 1991, 1999; Krosnick & Alwin, 1988), assigning similar ratings to items using the same scale. *Acquiescence bias* describes the tendency for satisficing participants to simply agree or disagree with statements regardless of their content (Moum, 1988; Winkler, Kanouse, & Ware, 1982). Satisficers also tend to respond more often with "don't know" (DKing) when such a response is offered (Schuman & Presser, 1981), and satisficers show *mark all effects*, selecting less items when questions ask respondents to "mark all items that apply" vs inquire about the relevance of every item individually (Smyth, Dillman, Christian, & Stern, 2006).

The extent to which participants satisfice varies with aspects of survey design. For example, longer surveys, which fatigue respondents, are more prone to satisficing behaviors (Krosnick & Alwin, 1988), and surveys on trivial or non-personally relevant topics, which participants spend less time thinking about, are also prone to satisficing (Krosnick, 1991; Holbrook, Green, & Krosnick, 2003; Holbrook, Krosnick, Moore, & Tourangeau, 2007). Satisficing also increases when questions are difficult to answer (Gage, Leavitt, & Stone, 1957). In addition, satisficing varies with individual difference variables, and satisficers have been found to be less intelligent and less politically informed (Holbrook et al., 2007; Krosnick & Alvin, 1988;

Narayan & Krosnick, 1996). Finally, the IMC development literature also suggests that satisficers are more likely to fail an IMC (Oppenheimer, et al., 2009).

Satisficing is conceptualized as existing on a continuum rather than being a dichotomous measure of present vs absent (Krosnick, 1991). Thus, participants may pass an IMC while still displaying some level of satisficing (Berinsky et al., 2013). Whereas previous research used IMCs as measures of attention, the present research asks whether exposure to an IMC is itself a treatment that influences how much attention respondents pay to subsequent questions. Do respondents show less satisficing after (than before) encountering an IMC question?

3 Conversational Effects

In everyday life, conversations follow a cooperation principle (Grice, 1975) that allows listeners to assume that speakers attempt to be informative, relevant, and clear. When speakers fail to live up to these expectations, listeners draw on the context of the utterance to infer its likely meaning (for reviews see Clark & Clark, 1977; Schwarz, 1994, 1996). Research participants bring these expectations to the research situation and consider all contributions of the researcher to be relevant to their task. These contributions include formal features of questionnaire design, from scale format to graphics and question wording. As a result, many "technical" aspects of questionnaires become a source of information that respondents systematically use to determine what is asked of them (for reviews, see Conrad, Schober, & Schwarz, 2014; Schwarz, 1994, 1996).

For instance, respondents draw on the numeric values of rating scales to interpret the intended meaning of verbal labels (Schwarz, Knäuper, Hippler, Noelle-Neumann, & Clark, 1991; Schwarz, Grayson, & Knauper, 1998), resulting in *scale value effects***.** They also assume that values in the middle range of a frequency scale reflect the population average (Schwarz, Hippler, Deutsch, & Strack, 1985), resulting in *scale range effects.* When encountering an ambiguous question, they draw on the content of prior questions to interpret its meaning, resulting in *question context effects* (Strack, Schwarz, & Wänke, Study 1, 1991). Throughout, respondents assume that the researcher is a cooperative communicator whose contributions are relevant to their task, consistent with the tacit assumptions underlying conversational conduct in everyday life (Schwarz, 1996). Accordingly, they pay close attention to subtle contextual features, in particular when they encounter ambiguous questions. The experience that the researcher presents a "trick" question may influence the emergence of Gricean conversational effects in different ways. On the one hand, learning that attention is called for may increase attention and hence the impact of subtle contextual cues; on the other hand, realizing that the researcher is not always a cooperative communicator may undermine reliance on conversational

norms and hence attenuate the influence of conversational inferences. Next, we turn to these potential influences.

4 Implications of IMCs for Survey Research

If IMCs alert participants that a question may not be what it seems at first glance (as shown by Hauser & Schwarz, 2015a), they may influence responses in a variety of ways. First, they may increase attention to ensure that one isn't "tricked" in subsequent questions. Second, they may teach respondents that the researcher is not a fully cooperative communicator, which may undermine respondents' reliance on conversational norms in making sense of the questions asked. These two possibilities result in differential predictions.

Increased attention

In survey questionnaires, increased attention to the details at hand should attenuate satisficing effects (Krosnick, 1991, 1999), that is, response effects that are commonly attributed to low attention and mental short cuts. The more attention respondents pay to the questions, the less they should resort to "top-of-the-head" answers. In contrast, increased attention to the details at hand should increase conversational inference effects, that is, response effects that are commonly attributed to the operation of conversational norms (Schwarz, 1994, 1995, 1996). These effects require close attention to minor question details (such as numerical values or scale range) in drawing inferences about a question's intended meaning; they should therefore benefit from increased attention. Note that these considerations entail that increased attention and effort have opposite effects on the emergence of satisficing and Gricean norm effects.

Cooperativeness

Complicating predictions, answering an IMC may also teach respondents that the researcher is not a fully cooperative communicator. Asking a question that seems to inquire about X, while noting along the way that X should be ignored in favor of a substantively unrelated response, violates the norms of cooperative conversational conduct (Grice, 1975). The impression that the researcher is not a cooperative communicator, in turn, may reduce the likelihood that participants draw on other features of the questionnaire to infer what the researcher may have had in mind (Schwarz, 1996). If so, response effects based on Gricean conversational processes should be attenuated (rather than increased) when the respective question is preceded by an IMC.

Motivation

Finally, being asked an IMC may also undermine respondents' motivation and willingness to live up to their role – they didn't agree to being "tricked", after all. If so, it may result in more missing data, early termination of online surveys, and so on. Our studies are not suited to assess this possibility because they draw on Amazon Mechanical Turk (MTurk) workers as participants. These online participants are paid for good performance and rely on positive ratings from their employers, which are the basis of reputation scores that drive their future employment. Accordingly, a transparent lack of cooperation is unlikely to be observed in samples of MTurk workers (see Hauser & Schwarz, 2015b).

Manipulation versus measure

Note that our analysis of IMCs treats IMCs as a manipulation of attention, not merely a measure of attention. Our predictions therefore deviate from the more familiar prediction that those who pass an IMC will show less satisficing than those who fail an IMC. The latter prediction pertains to an individual difference in attention and/or motivation and uses IMCs as a measure. In MTurk samples, more than 90 percent of participants routinely pass IMCs (see Hauser & Schwarz, 2015b), indicating that the situational incentives provided by performance-dependent payment and reputation ratings trump variations at the individual difference level. Thus, in the studies that follow, we restrict our analyses to only the participants who pass the IMC in order to assess its potential as a manipulation of attention.

5 Replication, Logic of Analysis, and Data Collection

We test whether previously documented changes in information processing moderate the emergence of context effects in surveys by presenting an IMC either before or after classic survey context effects. This design incorporates replications of classic effects into our investigation. We expect effects driven by satisficing and Gricean norms to replicate when such items precede an IMC, and we test predictions about how an IMC may affect their emergence and size when these items follow an IMC. Note that testing the effect of IMC order is mute if a classic effect does not replicate when administered before an IMC to begin with.

In two online surveys, we presented an IMC either before or after questions expected to elicit classic survey context effects. For ease of presentation, we discuss the satisficing and conversational experiments separately and note in which of the two surveys they appeared. The Method section that follows provides details regarding the online surveys.

6 Method

6.1 Survey 1

Participants

Seven hundred and ninety-eight American Amazon Mechanical Turk (MTurk) workers (456 male, age range 18 - 81) completed a survey in exchange for 40 cents. An a priori power analysis suggested this sample size yields an estimated 98% power for finding an effect of IMC order on satisficing measures when $d = .30$ for the effect of IMC order (Faul, Erdfelder, Lang, & Buchner, 2007).

Materials and procedure

Participants were directed to an online Qualtrics survey ostensibly on current issues. After consenting to the research, participants completed a battery of eight tasks and an IMC. Crucially, random assignment determined the order in which the task battery and IMC were administered. In one condition (IMC first), participants completed the IMC first, followed by the task battery. In the other condition (IMC last), participants completed the task battery, then the IMC. See Appendix A for wording of all questions.

Instructional manipulation check

The IMC was a standard attention check (adapted from Oppenheimer et al., 2009) which has been shown to affect systematic thinking in prior research (Hauser & Schwarz, 2015a) and which has been used extensively in unsupervised online research. In this question, a lure prompt asks participants to choose which of a long list of sports activities they regularly engage in, asking them to check all sports that apply. However, an instruction block informs participants that researchers are interested in their attention levels and, in order to demonstrate attention to the instructions, participants should only select the "other" option below and type in to the accompanying textbox "I read the instructions." Participants who followed these instructions were scored as "passing" the trap question.

Task battery

A battery of eight tasks assessed the degree to which participants exhibited survey context effects. Participants were randomly assigned to receive the tasks in different orders.

Question context and a fictitious issue

In an effort to cooperatively answer questions, participants often assume adjacent questions are related and use prior questions to draw inferences about ambiguous concepts. Modeled on Strack, Schwarz, and Wänke (1991), participants reported whether they favored or opposed (forced choice) a fictitious "Data Sharing Act".

This question was preceded by a question that either referred to Google's decision to grant users control over their personal data or to the U.S. governments' mass collection of private emails and browsing histories; these questions are predicted to provide a positive vs. negative context for interpreting what the fictitious Data Sharing Act refers to, resulting in differential support. These questions constitute a novel conceptual replication of previous experiments on fictitious issues.

Response order

Taken from Schuman & Presser (1981), two tasks assessed satisficing-driven response order effects**.** People taking mental shortcuts don't give full consideration to all response options and tend to select the first reasonable response they consider. When response options are presented visually, the first option is the first considered and is more often selected (Krosnick & Alwin, 1987; Schwarz, Strack, Hippler, & Bishop, 1991). Participants reported which of two statements they agreed with regarding the world's oil supply ("we will still have plenty of oil 25 years from now" or "it will all be used up in about 15 years") and the government's role in supplying adequate housing ("the federal government should see to it that all people have adequate housing" or "each person should provide for his own housing"); the order of responses options was manipulated.

Nondifferentiation

When faced with rating many items on the same scale, satisficers tend to assign many items the same rating. Modeled on Krosnick and Alwin (1988), in a single question matrix, participants rated their interest in thirteen reality television shows on a five point scale (1 = extremely interesting, 2 = very interesting, 3 = fairly interesting, 4 = not too interesting, 5 = not interesting at all). To compute nondifferentiation scores, we counted the number of shows to which participants assigned the same rating.

Don't know

Satisficers are more likely to give "don't know" (DK) responses when these options are offered as it is an easy response. Questions taken from Schuman and Presser (1981) asked about the severity of local courts and about federal government power, and participants were either offered a DK response option or not. All participants typed their response into textboxes, which we coded as falling into the various response options or as expressing a DK response.

Mark all effects

When asked to "mark all items that apply," satisficers tend to consider and mark only a few of the items. This results in less items selected compared to a question that forces respondents to consider each option individually. Modeled on Smyth, Dillman, Christian, and Stern (2006), participants indicated from which of 16 Amazon.com departments they had purchased items in the last 18 months. Partici-

pants were randomly assigned to either "mark all departments that apply" or were asked about each department individually.

Acquiescence
Satisficers often agree or disagree with a majority of statements and contradict themselves in their answers. Modified from Winkler, Kanouse, and Ware (1982), participants selected whether they agreed or disagreed with twenty statements concerning doctors and healthcare. Five pairs of statements (ten statements in total) were logical opposites, which assessed acquiescence bias. The remaining ten statements were filler items.

Task order
We varied the order in which the eight tasks were presented in order to a) assess whether the effects of the IMC on subsequent tasks vary as a function of distance from the IMC and b) assess the sensitivity of our measures to satisficing. We were interested in whether the effects of the IMC "wore off" and became less strong as an item was moved further away from the IMC. Half of the participants received the tasks in the following order: data sharing act, oil supply, reality TV shows, court punishment, adequate housing, Amazon purchasing, government power, and healthcare attitudes. The other half received the tasks in this order: data sharing act, adequate housing, Amazon purchasing, government power, oil supply, reality TV shows, court punishment, healthcare attitudes.

6.2 Survey 2

Participants

Three hundred and ninety seven participants from MTurk participated in the study (254 male, 143 female) in exchange for 40 cents. An a priori power analysis showed that when $d = .30$ (a conservative estimate of the effect size of IMC order) this sample size has 82% power for finding an effect of IMC order (Faul, et al., 2007). The median time to complete the survey was two minutes. We excluded the data of one participant who took twenty-seven minutes (nearly twelve standard deviations beyond the mean survey completion time) to complete the survey, bringing our total number of participants down to 396.

Materials and procedure

Participants were directed to a survey ostensibly addressing current issues. Participants completed an IMC and a series of Gricean conversational norm tasks. They were randomly assigned to receive the IMC as either the first or last question in the survey.

Instructional manipulation check

The IMC (adopted from Oppenheimer et al., 2009) followed the same format as in Study 1. However, unlike Study 1, participants were also randomly assigned to receive feedback on their response. Feedback informed participants of incorrect answers on the trap question and returned them back to the IMC with the instructions "Please try again" in the event of an incorrect response. Participants assigned to receive no feedback were not informed of incorrect answers, and thus simply progressed to the next page of the survey in the event of an incorrect response. However, because we restricted our analyses to only the participants who answered the IMC correctly (as detailed in the upcoming results section), none of our participants whose data was analyzed actually received feedback. Therefore, this manipulation was not included in our analyses and won't be discussed further.

Task battery

Participants completed three tasks designed to measure context effects due to inferences from conversational norms. The wording of all tasks is shown in Appendix A.

Scale range effects

Participants view scale ranges presented by researchers as being informative inputs for their judgments, assuming that middle values in the range reflect population averages. When asked how many hours of television they watch per day, participants given scales that contain more values below the population average (low-skewed scales) report watching less hours of television than participants given scales that contain more values above the population average (high-skewed scales). Additionally, when asked how important a role TV plays in their leisure time, participants given low-skewed scales report a more important role of TV than participants given high-skewed scales. Because participants given low-skewed frequency scales often rate their TV watching frequency above the scale's midpoint, this prompts them to infer that they watch more TV than average and think that TV plays a rather important role in their leisure time (and vice versa for high-skewed frequency scales; Schwarz, Hippler, Deutsch, & Strack, 1985).

Adapted from Schwarz et al. (1985), participants rated how many hours of TV they watch daily. Participants were randomly assigned to either a low frequency scale (ranging from "up to .5 hour" to "more than 4.5 hours") or a high frequency scale (ranging from "up to 4.5 hours" to "more than 8.5 hours"). The scale was created around the actual mean hours of TV viewed per day in America (4.5 hours; Nielsen, 2011), and both scale range conditions contained that mean. Following this question, participants were then asked, "How important is the role of TV in your leisure time?" with an 11-point scale (0 = "not at all important" to 10 = "very important").

Scale label effects

Participants draw on the numeric values of rating scales to infer question meaning. When asked how successful they have been in life, respondents report higher success when the scale runs from -5 ("not at all successful") to +5 ("extremely successful") rather than from 0 ("not at all successful") to 10 ("extremely successful"). This reflects that the bipolar -5 to +5 format suggests an interpretation that spans the whole range from failure (-5) to success (+5), whereas the unipolar 0 to 10 format covers only differential degrees of success (Schwarz, Knäuper, Hippler, Noelle-Neumann, & Clark, 1991). We replicated this experiment.

Similarly, participants provide higher ratings of the frequency with which they engage in rare behaviors when the rating scale runs from 0 ("rarely") to 10 ("often") rather than 1 ("rarely") to 11 ("often"). This is the case because "rarely" is interpreted as "never" when combined with 0 and interpreted as a small frequency when combined with 1, resulting in corresponding shifts on the scale (Schwarz, Grayson, & Knauper, 1998). We replicated this experiment with questions about the frequency of getting a haircut, visiting a museum, and attending a poetry reading.

7 Results

IMC performance

In survey 1, 747 participants (93.5%) answered the IMC correctly, while only 52 participants (6.5%) answered it incorrectly. This high IMC pass rate is consistent those of recent research on MTurk (Hauser & Schwarz, 2015b; Nauts, Langner, Huijsmans, Vonk, & Wigboldus, 2014; Wolf, Levordashka, Ruff, Kraaijeveld, Lueckmann, & Williams, 2014). Following convention (Oppenheimer et al., 2009) we restricted our survey 1 data to the sample of participants who answered the IMC correctly because this is the primary sample of interest. Moreover, the small number of participants who failed the IMC does not allow for meaningful comparisons of the question effects of interest.

In survey 2, 369 participants (93%) answered the IMC correctly on their first try. As with survey 1, we restricted our survey 2 sample to the 369 (93%) participants who responded correctly to the IMC because the sample of participants who responded incorrectly was not large enough for drawing firm conclusions.

Satisficing effects

For each question experiment, we first present replication analyses that assess whether the standard satisficing effect emerges when the IMC is the last task in the sequence, that is, under normal survey conditions without a potential IMC intervention. Subsequently, we test whether an observed effect is attenuated when the

Table 1 Summary of satisficing effect results

Satisficing-driven survey context effect	Replicates?	Moderated by IMC order?
response order effects (Schuman & Presser, 1981)		
oil supply	no	no
adequate housing	yes	no
nondifferentiation (Krosnick & Alwin, 1988)		
reality TV shows	yes	no
DKing (Schuman & Presser, 1981)		
court punishment	yes	no
government power	yes	no
mark all effects (Rasinksi et al., 1994)		
Amazon purchasing	yes	no
acquiescence (Winkler et al., 1982)		
healthcare attitudes	–	no

IMC precedes rather than follows the items of interest. Table 1 summarizes the conclusions.

Response order effects

One of the two response order questions in survey 1 asked whether the government should provide adequate housing (taken from Schuman & Presser, 1981). When the IMC was presented last, participants were more likely to choose "the government" as their response when it was the first response option listed (49%) than when it was the last option listed (36%); $\chi^2(N = 372, 1) = 5.74$, $p = .017$, $\varphi = .12$. This replicates to the standard response order effect.

To assess if prior IMC administration attenuated this effect, we conducted a logistic regression with IMC order (IMC first, IMC last), response option order (government first, government last), task order (2^{nd} task, 5^{th} task), and their interactions entered as mean-centered categorical predictors of response to the adequate housing question (1 = government, 2 = each person). Importantly, this response order effect was unaffected by prior answering an IMC, $\beta = -0.04$, $Wald = 0.35$, $p = .56$ for the 2 way interaction of IMC order and response order. As suggested by the replication analysis, the main effect of response option order was significant, $\beta = .30$, $Wald = 16.12$, $p > .001$, $OR = 1.35$. All other main effects and interactions failed to reach significance, ps $> .12$. In sum, prior exposure to an IMC did not attenuate the classic response order effect on this task.

A second response order question in survey 1 pertained to the *oil supply* (Schuman & Presser, 1981). Under standard conditions (IMC last) the familiar response order effect did not replicate, χ^2 (N = 372, 1) = 1.11, p = .29. Hence, this item cannot serve as an index of satisficing in our sample. (For additional analyses of this item see Appendix B.)

Nondifferentiation

One question in survey 1 concerning interest in reality TV shows assessed non-differentiation behavior. Survey fatigue effects suggest nondifferentiation should increase when the task is administered later in the survey (Krosnick, 1991). We replicated this effect when the IMC was presented last; the mean number of identically-rated shows was higher when the reality TV show question was presented sixth in the task battery (M = 9.38, SD = 2.80) compared to when it was presented third in the battery (M = 8.53, SD = 2.63); F(1, 369) = 9.01, p = .003, η_p^2 = .024, 95% CI [-1.40, -0.29] for the effect of task order.

To test for a potential effect of IMC placement, we conducted a 2 (IMC order: IMC first, IMC last) x 2 (task order: 3rd task, 6th task) between subjects analysis of variance on the number of shows given an identical rating. First answering an IMC did not affect nondifferentiation; F < 1 for the main effect of IMC order. The interaction of IMC order and task order also did not reach significance; F(1, 741) = 1.91, p = .168. As shown in the replication analysis, the main effect of task order was significant, F(1, 741) = 8.15, p = .004, η_p^2 = .011, 95% CI [-0.48, -0.09]. Thus, prior exposure to an IMC did not lessen participants' nondifferentiation behavior.

DK effects

Two questions in survey 1 assessed the influence of offering a DK option. When the IMC was presented last, the standard effect replicated for both questions. On the question regarding court punishment (Schuman & Presser, 1981), participants were much more likely to indicate a "don't know" response when a DK option was explicitly offered (58.1%) than when it was not explicitly offered (0%); χ^2(N = 373, 1) = 152.83, p < .001, φ = .64 for the effect of DK option.

In order to assess whether the experimental treatments significantly affected DK responses to this question, we limited our sample to the participants who were offered a DK option and conducted a logistic regression with IMC order, task order (4th task, 7th task), and their interaction entered as mean centered categorical predictors of giving a DK response (0 = non-DK, 1 = DK). Task order did not affect DK responses, β = -.17, *Wald* = 2.69, p = .101, *OR* = 0.84 for the main effect of task order. Prior answering an IMC also did not affect DK responses, β = .02, *Wald* = 0.07, p = .813 for the main effect of task order. The interaction of task order by IMC order was also not significant, β = .09, *Wald* = .78, p = .377. Thus, while standard DK effects replicated, prior exposure to an IMC did not significantly lessen the extent to which participants selected a DK response.

On the question regarding government power (Schuman & Presser, 1981), participants were again more likely to indicate a "don't know" response when a DK option was offered (14.0%) than when it was not (0%); $\chi^2(N = 372, 1) = 27.95$, $p < .001$, $\varphi = .27$.

In order to assess whether the experimental treatments significantly affected DK responses to this question, we limited our sample to the participants who were offered a DK option and conducted a logistic regression with IMC order, task order (4th task, 7th task), and their interaction entered as mean centered categorical predictors of giving a DK response (0 = non-DK, 1 = DK). DK responses were no more likely in either task order; $\beta = -.05$, *Wald* = 0.10, $p = .748$, for the main effect of task order. DK responses were also not affected by prior seeing an IMC; $\beta = -.01$, *Wald* = 0.00, $p = .968$ for the main effect of IMC order. Finally, the interaction of IMC order and task order was not significant, $\beta = -.23$, *Wald* = 2.27, $p = .131$. Thus, while standard DK effects replicated, prior exposure to an IMC did not significantly alter the extent to which participants selected a DK response for either question.

Mark all effects

One question in survey 1 regarding Amazon.com department purchases assessed mark all effects. Participants tend to select fewer options when given a mark all question type than when asked about each option individually (Smyth et al., 2006). We replicated this effect when the IMC was presented last; participants selected less departments when asked to *mark all* ($M = 3.7$, $SD = 2.5$) than when asked about each department separately ($M = 4.7$, $SD = 3.3$); $F(1, 370) = 11.69$, $p = .001$, $\eta_p^2 = .03$, 95% CI [-1.64, -0.44].

In order to assess if prior answering an IMC attenuates this effect, we conducted a 2 (IMC order: IMC first, IMC last) x 2 (task order: 3rd task, 6th task) x 2 (question type: mark all, individual questions) between subjects analysis of variance on the number of Amazon.com departments selected. As shown in the replication analysis, the main effect of question type was significant, $F(1, 738) = 22.60$, $p < .001$, $\eta_p^2 = .03$, 95% CI [-0.70, -0.29]. The effect of question type was also marginally moderated by task order: interaction of task order x question type, $F(1, 738) = 3.40$, $p = .066$, $\eta_p^2 = .01$, 95% CI [-0.01, 0.39]. Simple effects tests showed that when the task appeared as the 3rd task in the battery, participants selected less departments when given a *mark all* item type ($M = 3.5$, $SD = 2.3$) than when given an *individual questions* item type ($M = 4.9$, $SD = 3.3$); $F(1, 738) = 21.94$, $p < .001$, $r = .17$ for the simple main effect. When the task appeared as the 6th task in the battery, the effect of question type was in the same direction but less strong. In these conditions, participants selected less departments when given a *mark all* item type ($M = 3.6$, $SD = 2.4$) than when given an *individual questions* item type ($M = 4.2$, $SD = 3.2$); $F(1, 738) = 4.2$, $p = .041$, $r = .07$ for the simple main effect.

Importantly, the effect of question type was not attenuated by prior answering an IMC: $F < 1$ for the interaction of question type and IMC order. All other interac-

tions and main effects failed to reach significance, *p*s > .20. Thus, prior exposure to an IMC did not lessen classic "mark all" effects.

Acquiescence

Survey 1 also included an empirically-validated acquiescence scale that assesses how many contradictory statements regarding healthcare that a respondent endorses (Winkler et al., 1982). Prior exposure to an IMC did not lessen acquiescence on this scale, $F < 1$ for the effect of IMC order on the number of contradictory statement pairs each participant selected.

Gricean conversational norm effects

Next, we turn to Gricean conversational norm effects. For each experiment, we again report whether the original effect replicated and then assess whether its emergence and size is moderated by the placement of an IMC. Table 2 summarizes the analyses.

Question context and a fictitious issue

One question in survey 1 assessed whether participants used a preceding context question to disambiguate the meaning of a fictitious Data Sharing Act. Replicating the findings of Strack, Schwarz, and Wänke (1991) with a novel question set, when the IMC was presented last, a favorable context prompted more "favor" responses to the fictitious issue (46.5% favor) than an unfavorable context (9.1% favor) χ^2 (1, N = 372) = 34.95, $p < .001$, $\varphi = .42$.

In order to assess if this effect was moderated by IMC order, we conducted a logistic regression with IMC order (IMC first, IMC last), prior question context (favorable, unfavorable), and their interaction entered as mean-centered categorical predictors of approval of the fictitious issue (1 = favor, 2 = oppose). Consistent with the replication analysis, the main effect of prior question context was significant, β = .93, *Wald* = 89.43, $p < .001$, *odds ratio [OR]* = 2.53. All other effects failed to reach significance; β = .07, *Wald* = .45, $p = .50$, for the main effect of IMC order and β = .15, *Wald* = 2.38, $p = .12$, for the interaction of IMC order and context. Thus, placement of the IMC did not influence the extent to which participants drew on question context in interpreting an ambiguous issue.

Scale range effects – behavioral report

One question in survey 2 assessed whether reports of TV consumption were affected by the range of the frequency scale. Today, the average TV consumption in the United States is about 4.5 hours (Nielsen, 2011). When the IMC was presented last, 19.6% of the participants reported watching more than 4.5 hours when given the high frequency scale, whereas only 3.4% did so when given the low frequency scale; χ^2 (1, $N = 369$) = 11.47, $p = .001$, $\varphi = .25$. This replicates the original pattern reported by Schwarz et al. (1985) with values that have been adjusted to reflect current TV consumption.

Table 2 Summary of Gricean norm effect results

Gricean-driven context effect	Replicates?	Moderated by IMC order?
Question context and a fictitious issue (Strack et al., 1991)		
data sharing act	yes	no
Scale labels (Schwarz et al., 1991; Schwarz et al., 1998)		
life success	yes	no
rare behavior frequency	no	no
Scale range (Schwarz et al., 1985)		
TV consumption – behavioral report	yes	no
TV consumption – comparative judgment	yes	yes

To test if scale range effects are moderated by prior exposure to an IMC, we conducted a logistic regression with IMC order (first, last), scale range (low, high), and their interaction entered simultaneously as mean-centered categorical predictors of the likelihood of participants saying they watch more than the mean amount of TV per day (0 = no, 1 = yes). Importantly, IMC order did not moderate scale range effects, $\beta = 0.53$, $Wald = 0.40$, $p = .527$ for the two way interaction of IMC order and scale range. Consistent with the replication analysis above, the effect of the scale range was significant, $\beta = 1.66$, $Wald = 15.95$, $p < .001$, $OR = 5.28$ for the main effect. The main effect of IMC order was not significant, $\beta = -0.25$, $Wald = 0.38$, $p = .540$. Thus, IMC order did not affect this Gricean norm effect.

Scale range effects – comparative judgment

A follow-up question in survey 2 assessed whether judgments of TV's importance in participant's leisure activities were affected by the frequency scale presented with the behavioral question. Participants who report their behavioral frequency along a low (high) frequency scale endorse values in the higher (lower) range of the respective scale. As observed in previous research (Schwarz et al., 1985), participants infer their likely placement in the distribution from their placement on the scale. Hence, a low frequency scale suggests that their own TV consumption is above average, whereas a high frequency scale suggests that it is below average. This, in turn, affects judgments of how important TV is in their own lives. Replicating this effect, participants given a low frequency scale range rated TV as being more important to their leisure time ($M = 5.38$, $SD = 2.41$) than participants given a high scale range ($M = 4.62$, $SD = 2.63$); $F(1, 187) = 4.26$, $p = .040$, $\eta_p^2 = .02$, 95% CI [0.03, 1.49] for the effect of scale range when the IMC is presented last.

In order to investigate if IMC order moderates this effect, we conducted a 2 (IMC order: first, last) x 2 (scale range: low, high) between subjects analysis of variance on the importance of TV in participants' lives. There were no main effects, *p*s > .10. However, IMC order did marginally moderate the effect of scale range: $F(1, 361) = 3.18$, $p = .075$, $\eta_p^2 = .01$, 95% CI [-0.49, 0.02] for the interaction of IMC order and scale range. As shown before, when participants received the IMC last, there was the typical effect of scale range; those participants presented with a low scale range reported TV as being more important in their lives compared to those participants who received the high scale range: $F(1, 365) = 4.26$, $p = .040$, $r = .11$, 95% CI [0.02, 0.74] for the simple effect of scale range. However, this effect was eliminated when participants answered the IMC first. In this case, TV importance ratings did not differ ($M = 4.63$ and 4.80, $SD = 2.57$ and 2.49 for the low and high frequency conditions, respectively), $F < 1$ for the simple effect of scale range. Thus, IMC order moderated this effect. We discuss the implications of this observation in the General Discussion.

Scale label effects

Two tasks in survey 2 assessed scale label effects. When asked about their success in life, participants provide more modest ratings when the numeric values of the rating scale suggest that the low anchor of the scale refers to the absence of outstanding achievements (0 = not at all successful to 10 = very successful) rather than the presence of explicit failure (-5 = not at all successful to +5 = very successful; Schwarz et al., 1991). Replicating this effect, 44.7% of the participants endorsed a value in the lower half of the 0-to-10 scale, whereas only 30.5% of the participants did so on the -5 to +5 scale; χ^2 (1, $N = 189$) = 4.04, $p = .045$, $\varphi = .15$ for the effect of scale values when the IMC was asked last.

To assess if IMC order moderates this effect, we conducted a logistic regression, where IMC order (first, last), scale label numeric values (-5 to +5, 0 to 10), and their interaction were entered simultaneously as mean-centered categorical predictors of participants' placing themselves in the lower half of the respective life success scale (0 = no, 1 = yes). IMC order did not moderate the impact of the numeric scale values, $\beta = 0.42$, *Wald* = 0.94, $p = .332$ for the two way interaction of scale label and IMC order. Consistent with the replication analysis, the main effect of scale labels was marginally significant, $\beta = 0.41$, *Wald* = 3.56, $p = .059$, *OR* = 1.50. The main effect of IMC order was also not significant, $\beta = -0.23$, *Wald* = 1.15, $p = .283$. Thus, IMC order does not increase this Gricean norm effect.

For the second scale label task, participants reported their frequency of engaging in rare behaviors. In previous research, participants interpreted the verbal end anchor "rarely" as "never" when it was paired with the numeric value 0, but not when paired with the numeric value 1. As a result of this shift in scale interpretation, they provided higher ratings along a 0 to 10 scale than along a 1 to 11 scale (Schwarz, Grayson, & Knäuper, 1998). This influence of numeric scale values was

not observed in our sample of participants receiving the IMC last, $F < 1$. This nonreplication renders the task unsuitable for exploring the potential influence of IMC order on Gricean task interpretations.

8 General Discussion

Instructional manipulation checks (IMCs) aim to identify research participants who pay little attention. These participants may introduce noise. Hence, identifying and excluding them has been found to increase data quality (Oppenheimer et al., 2009). However, cognitive research into the question-answering process highlights that every measurement is also a treatment (e.g., Nebel, Strack, & Schwarz, 1989; for a discussion, see Sudman, Bradburn, & Schwarz, 1996). If so, answering an IMC may influence participants' performance on subsequent tasks. Supporting this possibility, Hauser and Schwarz (2015a) found that participants performed better on reasoning tasks that required careful analytic reasoning when an IMC preceded rather than followed the task. This observation is potentially worrisome for survey researchers – although attention to survey tasks is generally desirable, inducing the sample to pay more attention to a task than the population ever may under natural conditions can result in erroneous population estimates.

As far as standard survey questions are concerned, the present findings indicate that there is less reason to worry than the Hauser and Schwarz (2015a) results suggested. In two online surveys with MTurk workers we administered twelve question experiments, seven pertaining to satisficing effects and five pertaining to Gricean norm effects. Two conclusions stand out. First, as shown in Tables 1 and 2, the classic response effects were highly robust and replicated well. The two exceptions were a nonreplication of a response order effect on Schuman and Presser's (1981) oil supply item and an influence of the numeric values of a rating scale on behavioral reports (Schwarz et al., 1998). There are no obvious reasons for these nonreplications and their cause is of limited interest for the present research, which requires the replication of response effects to assess their potential moderation through the placement of IMCs.

Second, and more important, the placement of IMCs did not affect the emergence, direction, or size of response effects (see Tables 1 and 2). The single exception is the observation that the range of a behavioral frequency scale influenced subsequent comparative judgments under standard conditions (replicating Schwarz et al., 1985), but not when an IMC preceded the question. Considered in isolation, this observation would be consistent with the assumption that IMCs undermine participants' belief that the researcher is a cooperative communicator. However, this interpretation is thwarted by the fact that a preceding IMC did not attenuate the influence of the scale manipulation on the behavioral report itself; nor did IMCs attenuate any of the other Gricean effects.

In combination, our findings are good news for survey methodologists. Although IMCs can influence how participants approach complex reasoning tasks (Hauser & Schwarz, 2015a), they seem unlikely to affect how they approach standard survey questions. We assume that the crucial difference is in the apparent nature of the task. Reasoning tasks of the type used by Hauser and Schwarz (2015a; taken from Frederick, 2005, and Toplak, West, & Stanovich, 2011) invite erroneous answers because the first answer that leaps to mind is objectively wrong, which more effortful systematic thinking elucidates. These tasks assess intuitive versus reflective thinking and were designed in such a way that a person must reflect in order to recognize that the initial intuitive answer is wrong. Thus, these questions require an element of error detection for correct answers and many people experience the questions as "tricky".

This is not the case for questions that give rise to satisficing effects and Gricean effects in survey research. These questions often ask people's opinions about issues or estimations of their own behaviors and are hardly perceived as "tricky." Further, these questions often lack a clearly right or wrong answer, and are thus unlikely to initiate error detection processes. Accordingly, questions relating to satisficing may not invite the same suspicion as complex reasoning tasks. If so, prior exposure to an IMC may only initiate systematic thinking on later *tricky-seeming* tasks that have objectively correct answers (which participants can check via systematic reasoning) while having no effects on other tasks. These conjectures await systematic testing.

References

Berinsky, A.J., Margolis, M.F., & Sances, M.W. (2013). Separating the shirkers from the workers? Making sure respondents pay attention on self-administered surveys. *American Journal of Political Science*, 1-15.

Buhrmester, M., Kwang, T., & Gosling, S.D. (2011). Amazon's Mechanical Turk a new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science*, *6*, 3-5.

Chandler, J., Mueller, P., & Paolacci, G. (2014). Nonnaïveté among Amazon Mechanical Turk workers: Consequences and solutions for behavioral researchers. *Behavior Research Methods*, *46*, 112-130.

Conrad, F.G., Schober, M.F., Schwarz, N. (2014). Pragmatic processes in survey interviewing. In T. Holtgraves (Ed.), *Handbook of language and social psychology* (pp. 420-437). Oxford, UK: Oxford University Press.

Clark, H.H., & Clark, E.V. (1977). *Psychology and language.* New York: Harcourt, Brace, Jovanovich.

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*, 175-191.

Frederick, S. (2005). Cognitive reflection and decision making. *Journal of Economic Perspectives*, 25-42.

Gage, N.L., Leavitt, G.S., & Stone, G.C. (1957). The psychological meaning of acquiescence set for authoritarianism. *The Journal of Abnormal and Social Psychology*, *55*, 98-103.

Goodman, J.K., Cryder, C.E., & Cheema, A. (2013). Data collection in a flat world: The strengths and weaknesses of Mechanical Turk samples. *Journal of Behavioral Decision Making*, *26*, 213-224.

Grice, H.P. (1975). Logic and conversation. In P. Cole & J.L. Morgan (Eds.), *Syntax and semantics: Vol. 3. Speech acts* (pp. 41-58). New York: Academic Press.

Hauser, D.J. & Schwarz, N. (2015a). It's a trap! Instructional manipulation checks prompt increased effort on "tricky" tasks. *SAGE Open*, *5*, 1-6. doi:10.1177/2158244015584617

Hauser, D.J. & Schwarz, N. (2015b). Attentive Turkers: MTurk participants perform better on attention checks than do subject pool participants. *Behavior Research Methods*. Advance online publication. doi:10.3758/s13428-015-0578-z

Holbrook, A.L., Green, M.C., & Krosnick, J.A. (2003). Telephone versus face-to-face interviewing of national probability samples with long questionnaires: Comparisons of respondent satisficing and social desirability response bias. *Public Opinion Quarterly*, *67*, 79-125.

Holbrook, A.L., Krosnick, J.A., Moore, D., & Tourangeau, R. (2007). Response order effects in dichotomous categorical questions presented orally: The impact of question and respondent attributes. *Public Opinion Quarterly*, *71*, 325-348.

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

Krosnick, J.A. (1999). Survey research. *Annual Review of Psychology*, *50*, 537-567.

Krosnick, J.A. & Alwin, D.F. (1987). An evaluation of a cognitive theory of response order effects in survey measurement. *Public Opinion Quarterly*, *51*, 201-19.

Krosnick, J.A., & Alwin, D.F. (1988). A test of the Form-Resistant Correlation Hypothesis Ratings, Rankings, and the Measurement of Values. *Public Opinion Quarterly*, *52*, 526-538.

Mason, W., & Suri, S. (2012). Conducting behavioral research on Amazon's Mechanical Turk. *Behavior Research Methods*, *44*, 1-23.

Moum, T. (1988). Yea-saying and mood-of-the-day effects in self-reported quality of life. *Social Indicators Research*, *20*, 117-139.

Nauts, S., Langner, O., Huijsmans, I., Vonk, R., & Wigboldus, D.H. (2014). Forming impressions of personality. *Social Psychology*, *45*, 153-163

Nielsen. (2011). *State of the media: The cross-platform report*. New York: The Nielsen Company.

Narayan, S., & Krosnick, J.A. (1996). Education moderates some response effects in attitude measurement. *Public Opinion Quarterly*, *60*, 58-88.

Nebel, A., Strack, F., & Schwarz, N. (1989). Tests als Treatment: Wie die psychologische Messung ihren Gegenstand verändert. [Tests as treatments.] *Diagnostica, 35*, 191-200.

Norenzayan, A., & Schwarz, N. (1999). Telling what they want to know: participants tailor causal attributions to researchers' interests. *European Journal of Social Psychology*, *29*, 1011-1020.

Oppenheimer, D.M., Meyvis, T., & Davidenko, N. (2009). Instructional manipulation checks: Detecting satisficing to increase statistical power. *Journal of Experimental Social Psychology*, *45*, 867-872.

Paolacci, G., Chandler, J., & Ipeirotis, P.G. (2010). Running experiments on amazon mechanical turk. *Judgment and Decision Making*, *5*, 411-419.

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

Schwarz, N. (1994). Judgment in a social context: Biases, shortcomings, and the logic of conversation. *Advances in Experimental Social Psychology, 26,* 123-162.

Schwarz, N. (1995). What respondents learn from questionnaires: The survey interview and the logic of conversation. (The 1993 Morris Hansen Lecture) *International Statistical Review, 63*, 153-177.

Schwarz, N. (1996). *Cognition and communication: Judgmental biases, research methods and the logic of conversation*. Hillsdale, NJ: Erlbaum.

Schwarz, N. (1999). Self-reports: how the questions shape the answers. *American Psychologist*, *54*, 93-105.

Schwarz, N., Grayson, C.E., & Knäuper, B. (1998). Formal features of rating scales and the interpretation of question meaning. *International Journal of Public Opinion Research, 10*, 177-183.

Schwarz, N., Hippler, H.J., Deutsch, B., & Strack, F. (1985). Response scales: Effects of category range on reported behavior and comparative judgments. *Public Opinion Quarterly*, *49*, 388-395.

Schwarz, N., Knäuper, B., Hippler, H.J., Noelle-Neumann, E., & Clark, L. (1991). Rating scales numeric values may change the meaning of scale labels. *Public Opinion Quarterly*, *55*, 570-582.

Schwarz, N., Strack, F., Hippler, H.J., & Bishop, G. (1991). The impact of administration mode on response effects in survey measurement. *Applied Cognitive Psychology*, *5*, 193-212.

Schwarz, N., Strack, F., & Mai, H.P. (1991). Assimilation and contrast effects in part-whole question sequences: A conversational logic analysis. *Public Opinion Quarterly*, *55*, 3-23.

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

Strack, F., Schwarz, N., & Wänke, M. (1991). Semantic and pragmatic aspects of context effects in social and psychological research. *Social Cognition*, *9*, 111-125.

Sudman, S., Bradburn, N.M., & Schwarz, N. (1996). *Thinking about answers: The application of cognitive processes to survey methodology*. Jossey-Bass.

Toplak, M.E., West, R.F., & Stanovich, K.E. (2011). The Cognitive Reflection Test as a predictor of performance on heuristics-and-biases tasks. *Memory & Cognition*, *39*, 1275-1289.

Winkler, J.D., Kanouse, D.E., & Ware, J.E. (1982). Controlling for acquiescence response set in scale development. *Journal of Applied Psychology*, *67*, 555-561.

Wolf, W., Levordashka, A., Ruff, J.R., Kraaijeveld, S., Lueckmann, J.M., & Williams, K.D. (2014). Ostracism Online: A social media ostracism paradigm. *Behavior Research Methods*, 1-13.

Appendix A

Survey 1 materials

Data sharing act

(favorable context) How do you feel about Google's decision to allow users complete control over the data they share (or choose not to) with advertisers?

(unfavorable context) How do you feel about the government's decision to allow government agencies to collect privately-shared data from internet users' email accounts and browsing histories?

Congress has been considering the Data Sharing Act of 2013. Do you favor or oppose the passage of this act?

Oil supply

Some people say that we will still have plenty of oil 25 years from now. Others say that at the rate we are using our oil, it will all be used up in about 15 years. Which of these ideas would you guess is most nearly right?

Adequate housing

Some people feel the federal government should see to it that all people have adequate housing, while others feel each person should provide for his own housing. Which comes closest to how you feel about this?

Reality TV shows

Please look at the reality television shows listed below. Could you please tell me whether you find the reality television show to be extremely interesting, very interesting, fairly interesting, not too interesting, or not interesting at all?

- The Real Teenagers of Beverly Hills
- Survivor
- Fish Tank Kings
- The Biggest Loser
- Hell's Kitchen
- So You Think You Can Dance?
- Shahs of Sunset
- Geeks vs. Greeks
- Married to a Vampire
- America's Next Top Model
- Millionaire Matchmaker
- The Bachelor
- The Apprentice

Court punishment

In general, do you think that the local courts in your area deal too harshly or not harshly enough with criminals (or do you not have enough information to say)? Enter "too harshly" or "not harshly enough" (or "not enough info") in the text box below.

Government power

Some people are afraid the government in Washington is getting too powerful for the good of the country and the individual person. Others feel that the government in Washington is not getting too strong. (Have you been interested enough in this to favor one side over the other? If so,) What is your feeling, do you think the government is getting too powerful or do you think the government is not getting too strong? Enter ("not interested enough,") "too powerful" or "not too strong" in the text box below.

Amazon purchasing

(mark all) From which of the following departments on Amazon.com have you made a purchase in the last eighteen months? (Check all that apply)

(individual questions) Have you or have you not purchased from the following departments on Amazon.com in the last eighteen months? (Select Yes or No)

- Unlimited Instant Videos
- MP3s and Cloud Player
- Amazon Cloud Drive
- Kindle
- Appstore for Android
- Digital Games and Software
- Audible Audiobooks
- Books
- Movies, Music & Games
- Electronics and Computers
- Home, Garden & Tools
- Grocery, Health & Beauty
- Toys, Kids & Baby
- Clothing, Shoes & Jewelry
- Sports & Outdoors
- Automotive & Industrial

Healthcare attitudes

Please look at the statements below and indicate whether you agree or disagree with each statement.

Doctors don't always explain to their patients the risks involved in certain treatments
(a) There is little a person can do to prevent illness
I'd rather my doctor just told me what to do
(b) Doctors do not always check everything they should check when examining their patients
Good doctors nearly always agree on how to treat a specific illness
(c) Prescription drugs frequently do more harm than good
Good health is largely a matter of luck
(d) Most doctors carefully explain what will happen to their patients
It mainly takes good medical care to get over an illness
Going to the doctor's office for check-ups is necessary
In the long run, people who take good care of themselves stay healthier and get well more quickly
(a) Anyone can learn a few basic health rules, which will go a long way in preventing illness
(e) A person should take medicine only as a last resort
It is important to seek immediate medical advice when you notice something wrong or unusual
(d) Doctors don't usually explain your medical problems to you
Sometimes doctors prescribe treatments that involve unnecessary risks
Your health is based more on genetics than the environment
(b) Doctors are very careful to check everything when examining their patients
(e) It's always silly to suffer if medicine will make you feel better
(c) Prescription drugs are almost always helpful

Survey 2 materials

TV consumption

On average, how many hours of TV do you watch daily?
(low frequency scale) Up to .5 hour, .5 hour to 1.5 hours, 1.5 hours to 2.5 hours, 2.5 hours to 3.5 hours, 3.5 hours to 4.5 hours, More than 4.5 hours
(high frequency scale) Up to 4.5 hours, 4.5 hours to 5.5 hours, 5.5 hours to 6.5 hours, 6.5 hours to 7.5 hours, 7.5 hours to 8.5 hours, More than 8.5 hours
How important is the role of TV in your leisure time?
1 = not at all important to 10 = very important

Life success

How successful have you been in life so far? Please use the following rating scale from -5 (not at all successful) to +5 (extremely successful) [from 0 (not at all successful) to 11 (extremely successful)].

Rare behavior frequency

How often do you get a haircut?
0 (1) = rarely to 10 (11) = often
How often do you visit a museum?
0 (1) = rarely to 10 (11) = often
How often do you attend a poetry reading?
0 (1) = rarely to 10 (11) = often

Appendix B

Table A1 Task order by IMC order by response order on *oil supply* response selection

	oil supply is 2nd question in battery				*oil supply* is 5th question in battery			
	IMC first		IMC last		IMC first		IMC last	
	plenty 1st	plenty 2nd	plenty 1st	plenty 2nd	plenty 1st	plenty 2nd	plenty 1st	plenty 2nd
plenty	58%	48%	58%	58%	47%	62%	59%	48%
used up	42%	52%	42%	42%	53%	38%	41%	52%

We conducted a logistic regression with IMC order (IMC first, IMC last), response option order (plenty first, plenty last), task order (2nd task, 5th task), and their interactions entered as mean-centered predictors of responses to the *oil supply* question (1 = plenty, 2 = used up). While the three way interaction of task order by IMC order by response option order was significant, $\beta = .18$, $Wald = 6.04$, $p = .014$, $OR = 1.20$, the patterns did not replicate the usual response order effect in any of the conditions (see Table A1) and is thus uninformative.

Information for Authors

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

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

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

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

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

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

Entire Book:

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

Journal Article (with DOI):

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

Journal Article (without DOI):

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

Chapter in an Edited Book:

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

Internet Source (without DOI):

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

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

gesis
Leibniz Institute for the Social Sciences

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