Effects of Question Characteristics on Item Non-response in Telephone and Web Survey Modes

Oliver Lipps[1] & Gian-Andrea Monsch[2]

Abstract:

Telephone surveys face more and more criticism because of decreasing coverage and increasing costs, and the risk of producing socially desirable answers. Consequently, survey administrators consider switching their surveys to the web mode, although the web mode is more suspectable to item non-response. Still, we do not know whether this is true for all question types. In this paper, we analyze to what extent item non-response depends on question characteristics such as their form or difficulty in the telephone and the web mode. We use data from an experiment in which individuals randomly sampled from a population register are experimentally assigned to these two modes. Distinguishing effects on the frequency of don't know responses, item refusals, and mid-scale responding, we find more don't know responses and item refusals for the web mode generally, but no differences for mid-scale responding. However, this relationship depends on the characteristics of the question.

Keywords: Mode effects, item non-response, question characteristics, telephone, web

Word count: 5,990

[1]Corresponding author, FORS, Lausanne, c/o University of Lausanne and University of Bern: oliver.lipps@fors.unil.ch
[2]Swiss Federal Statistical Office: gianandrea.monsch@bfs.admin.ch

Introduction

Item non-response is an important concern as it decreases statistical power if it occurs randomly and produces systematic bias if it is selective. While the web mode is becoming increasingly popular, this mode is more likely to produce item non-response compared to interviewer-based modes such as telephone surveys (de Leeuw 2005; Lesser, Newton and Yang 2012). One reason for this is the absence of an interviewer to motivate giving a valid answer, or to provide help if necessary (Bowling 2005; Groves et al. 2011). A second reason is that web surveys have an increased risk that respondents rush through the questionnaire or skip parts of a question, which may result in a reduced cognitive effort and a suboptimal answer, also termed "satisficing" (Simon 1957), instead of "optimizing" (Krosnick 1991). Both, no help from an interviewer and possible satisficing behavior in web surveys may increase the likelihood of giving an item non-response in the web mode, especially to more complex questions. However, research about the effects of the survey mode on item non-response depending on question characteristics is scarce if not completely absent from the literature (Revilla 2012). In this paper we go beyond effects of question difficulty, which have already been investigated in the literature, and add aspects such as form and content of questions. Thereby, we show that item non-response depends on both multiple aspects of question complexity and the survey mode.

We use experimental data from a pilot survey testing effects of telephone versus web on don't know responses, item refusals, and mid-scale responding. We distinguish questions with respect to their form, placement within the questionnaire, and complexity. After a literature review, we present our expectations of question characteristic effects in the two modes. We then describe the data and method we employ before discussing the results. We conclude with a discussion and present avenues for further research.

What influences Item (Non-)responses?

The literature mentions three main factors that affect item non-response: respondent characteristics, question characteristics, and survey mode. In terms of *respondent characteristics*, research has shown that age and education are related to item non-response. Older, very young, as well as lower educated people have more problems with question comprehension and/or answering (Krosnick, Narayan, and Smith 1996; Narayan and Krosnick 1996; Fricker, Galesic, Tourangeau, and Yan 2005; Yan and Tourangeau 2008; Couper and Kreuter 2013). Besides age and education, the comprehension of survey items also varies between languages and cultures (Warnecke et al. 1997; Harkness, Van de Vijver, and Mohler 2003). As Holbrook and colleagues suggest (2006), the way survey questions are constructed relies on cultural conceptions of the dominant cultural group and, therefore, respondents from other groups have more difficulties understanding survey questions. In this context, language as a cultural lens and language ability as such may also contribute to differences in item non-response (Johnson et al. 2006; Peytcheva 2008; Kleiner, Lipps, and Ferrez 2015). Finally, the motivation of respondents is related to item non-response (Cannell, Miller, and Oksenberg 1981; Kleiner et al. 2015) because motivation decreases satisficing (Krosnick 1991) and increases the cognitive effort of the interviewee when responding to survey questions (Edwards and Cantor 2004).

In addition, *question characteristics* affect item non-response. Especially complex questions are more difficult and take more time to respond to (Hope et al. 2014). This may result in a misunderstanding of questions, satisficing, or stopping the survey altogether. Questions become more difficult in content when they are subjective (Bassili 1996; Yan and Tourangeau 2008), abstract or vague (Holbrook, Cho, and Johnson 2006; Johnson et al. 2006), use unfamiliar words (Lenzner, Kaczmirek, and Lenzner 2009), require memory demands (Tourangeau, Rips,

and Rasinski 2000), are hypothetical (Lenzner et al. 2009), or include calculations and numeric values (Johnson et al. 2006; Lenzner et al. 2009). These characteristics increase the odds that respondents can't or do not want to answer.

In addition to their content, question forms can make questions harder for cognitive processing and responding. In particular, long questions (Holbrook et al. 2006; Johnson et al. 2006), questions with multiple clauses (Yan and Tourangeau 2008), complex syntax (Lenzner et al. 2009), as well as response framing such as acquiescence-prone questions (Saris et al. 2010) are known to have an influence on survey responding. Regarding response categories, the number and types may increase the difficulty to respond to a survey question (Tourangeau et al. 2000; Yan and Tourangeau 2008). Finally, the placement of questions within a questionnaire influences the cognitive effort of the response (Yan and Tourangeau 2008). Usually, effort decreases with survey progression.

While respondent and question characteristics are well established factors in the literature on item non-response, research on *mode effects* is relatively new. Current findings show that different modes produce different response patterns for different people and question characteristics. This may be due to people's different dispositions to giving socially desired answers or to satisfice, to the question order, to interviewer presence or absence, or to the visual layout of questions (Atkeson and Adams 2018; Fowler, Roman, and Di 1998; Leigh and Martin 1987; Schuman and Presser 1981; Schuman 1992; Sudman, Bradburn, and Schwarz 1996; Christian and Dillman 2004; Smyth et al. 2006; Tourangeau et al. 2004). In addition, effects of question characteristics may depend on the survey mode. While self-administered modes generally provide more missing data (Bowling 2005, de Leeuw 1992, Guzy and Leitgöb 2015), this is particularly true if questions are difficult or do not relate to their circumstances (Greene,

Speizer, and Wiitala 2008). Also, web surveys produce higher item non-responses for variables where a numeric value has to be entered, such as the value of one's property, the amount left to pay off on a mortgage, or weekly pay (Goodman et al. 2020; Voorpostel et al. 2020). Although these are rather sensitive questions and one might expect that the anonymity of the Web would produce less item non-responses, there seems to be no relation between sensitivity and item non-response (Tourangeau et al. 2000; Tourangeau and Yan 2007). That web respondents have more time and can reread the question may reduce cognitive burden particularly, when a question is difficult (Fricker et al. 2005), but this does not seem to decrease item non-response.

Survey Modes and Item Non-response

We define item non-response as a form of missing data where responses to some items are invalid (Groves et al. 2004; see also Young 2012). We measure three forms of item non-response: don't know (DK) responses, item refusals, and mid-scale responding. These measures have already been employed in other research on item non-response (Kleiner et al. 2015). All three of them are of course not exclusively item non-response. DK answers can be interpreted in various ways such as satisficing, passive refusal, random answer style, or valid responses (Young 2012). A similar argument can be made for mid-scale responding as well as item refusals. Nonetheless, we believe in the usefulness of such a distinction as they allow us to test whether survey modes have different effects on various types of item non-response. For example, we could expect more DK answers or refusal in self-administered questionnaires than in surveys administered by an interviewer and, consequently, a larger part of such answers become valid response in the latter

case. In contrast, the contrary can be expected for mid-scale responses where respondents give such answers to interviewers to hide item non-response (Sturgis, Roberts, and Smith 2014).

According to the literature, we expect survey modes to affect survey responding even controlled for respondent and question characteristics (Atkeson and Adams 2018). Specifically, we expect respondents surveyed by an interviewer to provide fewer DK answers or item refusals than respondents of a self-administered questionnaire. Even if one does not want or cannot answer a question, the presence of the interviewer exerts social pressure to give an answer. In contrast to DK answers and item refusal, we expect respondents surveyed by an interviewer to provide more mid-scale responses as this can be used as a technique to hide DK, item refusal, or indifference. In addition, we expect that mode effects on item non-response depend on question characteristics. The tendency of web respondents to give more DK answers and item refusals, and respondents administered by interviewers to give more mid-scale responses should increase with the content as well as the form of the question, such as the need of calculation, the length of a question, or the placement of a question within the questionnaire.

Data and Methods

We use data from the Swiss Household Panel (SHP; see Tillmann et al. 2016), more specifically from the pilot study for the SHP IV refreshment sample. After using the telephone as the main survey mode (for all respondents except for non-response follow-up and those without a landline), the SHP conducted a pilot study in fall/winter 2017/2018 to better understand how different mixed-mode designs may influence survey responding (details in Voorpostel et al. 2020). Questionnaires in the SHP include two household questionnaires: the grid (or household

roster, where all household members are enumerated and their basic socio-demographic characteristics captured) and the main household questionnaire (including characteristics about the dwelling, household finances, household task division). The household questionnaires are completed by the 'household reference person', who is an adult familiar with the financial situation of the household. After that, the individual questionnaires (multi-topic questionnaire about all aspects of individual living conditions) are completed by all household members from the age of 14 years on. The SHP is fielded in Swiss-German, French, and Italian, and respondents can choose the language.

The pilot study was an experiment using three groups, randomly assigned on the household level. The first group serves as a control group as they went through the traditional interview design of the SHP, which uses the telephone if available and face-to-face otherwise, and web as a non-response follow-up mode. Group 2 was interviewed by telephone on the household level if available and face-to-face otherwise and then pushed to a web survey for the individual questionnaire. Finally, group 3 undertook an entirely web-based survey on both levels. The web survey was a slightly adapted version of the telephone survey (modification of instructions, screen presentation) and could be completed with different devices (computer, smartphone, tablet, etc.)

The sample included 4'195 individuals living in private households in Switzerland drawn at random by the Swiss Federal Statistical Office. 790 individuals were assigned to Group 1 (478 telephone and 312 face-to-face on the household level), 2,192 to Group 2 (1,296 telephone and 896 face-to-face on the household level), and 1,213 to Group 3 (all web). The fieldwork was as similar as possible for members of the three groups. For example, all households received preliminary letters including a 10.- Swiss francs incentive per household member. All

nonresponding households received refusal conversion attempts. Individual response rates, meaning with the grid and the individual questionnaire completed, amounts to 37% for Group 1 (54% grid questionnaire and conditional 69% individual questionnaire), 36% for Group 2 (53% / 67%), and 33% for Group 3 (53% / 62%). The completion of individual questionnaires took on average about 53 minutes in the telephone mode, and about 56 minutes in the web mode. Excluding the face-to-face respondents resulted in 3,295 web or telephone respondents. In Group 1. 94 percent of individual interviews were conducted by telephone, in Group 2, 11 percent and in Group 3, 10 percent. We further excluded all respondents with no landline at home (variable included in the register; N=1,030) to make Group 1 respondents comparable with those from the other groups. Finally, we retained only those adult respondents who were eligible to answer all coded questions, so our analysis sample comprises 647 telephone and 1,467 web respondents.

We use the following independent variables for the respondent characteristics: survey language (French as the reference category, Swiss-German, Italian), survey language competence (survey language is first best-mastered language as the reference category, is second best-mastered language, is neither first nor second best-mastered language), nationality (Swiss as the reference category, from a bordering country, from another country), age groups (18-29 (reference), 30-39, 40-49, 50-59, 60-69, 70+ years old), highest education level (seven ordinal categories, modelled as a continuous variable), gender (reference female). In addition, we control for the status 'household reference person'. Since this person completed the two household questionnaires in addition to his/her own individual questionnaire, this may have had effects on survey responding in the individual questionnaire. In group 2, the grid/household questionnaires were completed via telephone, while individual surveys were conducted online. Consequently, we also control for a

mode change of the grid/household questionnaire of these respondents (reference: no mode change).

We use a subset of 209 questions of the 395 questions that were already asked in the SHP 2004 individual questionnaire and coded and analyzed in our previous analysis (see Kleiner et al. 2015 for details). These variables constitute most of the core questions of the SHP individual questionnaire and are effective in covering the range of topics and the characteristics of the questions asked in the SHP 2017 individual questionnaire.

Concerning the codings of question characteristics, we use three indicators for question content: memory, calculation, and type. We distinguished whether questions required use of long-term memory for something that occurred in the past: we distinguished between intensive searching, moderate searching, and little or no searching in long-term memory (reference category). Questions were also coded for whether they required some degree of calculation on the part of respondents (binary). Finally, we distinguished between three question types: evaluative (reference category), demographic, and factual. While demographic questions referred to social characteristics of respondents or related individuals, factual/behavioral questions referred to facts, conditions, or behaviors of or around respondents, and evaluative questions (attitudes/beliefs/expectations) involved some appeal to the judgment of respondents.

We further distinguished between three questions forms: question length, response format, and placement. A question was deemed long (binary) if it contained 20 and more words (done for German, French, and Italian separately). A first response format required a yes or no answer. Multiple response questions were those with three response options on one hand, and four or more on the other (reference category). Finally, the placement of questions in the survey, i.e., the

first third (reference category), second third, or last third was coded, since whether a question is earlier or later in a survey interview could influence response quality, given the effects of fatigue or declining motivation.

To evaluate the effects of question and respondent characteristics in our binary dependent variables (e.g., don't know vs. substantive answer given), we generated six binary logistic regression models (three dependent variables by mode) with the data transformed to a "cross-classified" data structure (see Fielding and Goldstein 2006). This structure allows for the consistent analysis of both respondent and question characteristics in the same model. The data set for the DK and item refusal cross-classified models consisted of 282,866 respondent-question combinations (83,296 telephone and 199,570 web), and the data set for the mid-scale model of 63 question variables and 92,670 respondent-question combinations (27,186 telephone and 65,490 web).

Effects of Mode, Respondents, and Question Characteristics

Table 1 presents the results of the six models with the two crossed levels respondents and question characteristics for DK answers, item refusal, and mid-scale responding by survey mode. First, we see that the constants for the web surveys models are higher than the respective constants for the telephone surveys for DK answers (Models 1 vs. 2) and for item refusal (Models 3 vs. 4) but not for mid-scale responding (Models 5 vs. 6). The confidence interval of the constants of the DK and the item refusal models do not overlap. This shows that when controlled for respondent and question characteristics, using the web mode produces significantly more DK answers (2.8% of all questions considered) and more item refusals (3.3%) compared to the telephone mode (1.4% and 0.3%). In contrast, the number of mid-scale responses for web

respondents (15.2%) and for interviews made via telephone (16.1%) are not statistically different, when controlled for respondent and question characteristics.

Table 1 about here

In addition to the net effect of modes, *question characteristics also matter for survey responding*. For the content of the question, it is mostly the strength of memory retrieval that is important. Examples of questions requiring higher memory retrieval include asking about the number of days the respondent was affected by health problems during the last year, about the number of years the respondent has resided in Switzerland or about the number of employees in his company. Such questions tend to produce in both modes more DK answers (Model 1 and 2) and mid-scale responding (Model 5 and 6).[3] Item refusal, in contrast, seems not to be affected by the strength of memory retrieval nor by the question content more generally. Questions needing calculation only increase DK answers in the web mode (Model 2). Finally, the type of question (evaluative, demographic, or factual) has no impact on item non-response.

Furthermore, question forms have more relevance for telephone than for web surveys. Fewer telephone respondents give DK answers when they were long, had fewer choice options and during the first third of the questionnaire (Model 1). Apart from the questionnaire length effect, we find the same pattern for item refusals by telephone (Model 3). In contrast, web respondents only give more DK answers during the last third (Model 2) and refuse items more often when there are three response options and during the last two-thirds of the questionnaire (Model 4).

---

[3] We only have two categories of memory retrieval (low and medium) for mid-scale responding, as none of the questions require a high level of memory retrieval.

Finally, question forms hardly affect mid-scale responding. Only questions in the last third of the questionnaire produce somewhat more mid-scale responses by web (Model 6).

Finally, *respondent characteristics* also have an impact on item non-response. In contrast to question characteristics, the type of respondent is more relevant for web respondents. Since respondents self-select into different modes to some extent (Voorpostel et al. 2020) these coefficients reflect both, self-selection and measurement effects. Because it is difficult to disentangle the two effects (Roberts and Vandenplas 2017), we only briefly describe the coefficients without interpreting them in the discussion. In addition, our main focus are effects from question characteristics and mode, so we treat respondent characteristics as controls rather than as substantively interesting covariates.

First, people, whose mother tongue is not one of the languages in which the questionnaire is administered, come from an adjacent country, who are less educated, and women give more DK answers, while Swiss-German speakers, and mid-aged people (40-59) give fewer DK answers by telephone (Model 1). In web surveys, Swiss-German speaking, older (from 30 on), better educated, and male web respondents give fewer DK answers while the contrary holds for Italian speakers, those interviewed in neither their first nor their second-best language, and foreigners from still another country (Model 2). Respondents who refuse single items in telephone surveys are less numerous among those people whose mother tongue is not one of the languages in which the questionnaire is administered, and more numerous among Italian speakers, foreigners, and the more educated (Model 3). Via web, Swiss-German speakers, younger people (up to 59), the more educated and men refuse less, while Italian speakers, whose mother tongue is not one of the languages in which the questionnaire is administered, and still other foreigners refuse more (Model 4). Reference persons refuse more by telephone. They give more DK answers and even

more item refusals by telephone but fewer by web if they must change mode between the household and the individual questionnaire. For mid-scale responding only age and education have an effect: older (from 70 on) telephone respondents give more mid-scale answers by telephone and older (from 30 onwards) and less educated more via web.

Discussion

Based on data from the SHP IV pilot study, which incorporated a mode experiment including web and telephone interviews, we explored the impact of different modes on item non-response for different question characteristics. Expecting a worse quality for web, we find that the web survey indeed produces more DK answers (delta=1.4 % points) and item refusals (delta=3.0 % points). Less mid-scale responding in the web (delta=0.9 % points) is not significant. This shows that – in line with the literature – web produces more item non-response than telephone, especially item refusals. Some of these item non-responses are likely to be hidden as mid-scale responses in the telephone mode. Although part of this difference is due to self-selection in the two modes, the differences are stable as the intercepts in multivariate models confirm.

Respondents who answer via telephone and web react differently to various question characteristics. High memory requirements produce more DK and more mid-scale responses in both the telephone and the web. Item refusals are not impacted by this kind of cognitive effort needed to answer a question. This suggests that respondents give honest answers (or a mid-scale answer) in case they really do not know the true answer, irrespective of the survey mode. Similarly, respondents give more DK answers if calculation is needed in the web mode. Since this is not the case in the telephone mode, this may suggest that interviewers are actually able to

help respondents to answer such questions. Interestingly, while long questions produced *fewer* DK in the telephone mode and did not affect web or the other measures, having fewer (two or three) response options produced both fewer DK and item refusals in the telephone mode and three response options slightly more item refusals in the web mode. It is likely that long questions contain long explanations read by the interviewers after which respondents may hesitate to give a DK. In contrast, having many response options may overburden respondents in the telephone setting such that they may not remember all the options. This is consistent with more recency effects in telephone surveys, i.e., the tendency to choose one of the later response options heard (e.g., de Leeuw 2005).

Finally, questions later in the questionnaire received more DK and item refusals in both modes and more mid-scale responses in the web mode. This may be a sign of fatigue, which is most pronounced by the much higher likelihood of giving a DK during the final third of the survey. It is also worth mentioning that the questionnaire lasts about an hour. With such a length, the impact of the placement may be relatively high. For item refusal, the difference between the coefficients of the middle and the final third of the survey is small which may indicate that, in particular, DK is sensitive to the length of a questionnaire.

Differential item non-response across modes and question characteristics may bias the results (e.g., Guzy and Leitgöb 2015). If there are differences between categories of certain person groups (such as between national languages, survey language competence, nationality, age, or gender), certain person groups will be underrepresented if – as is common in many analyses – missing data is listwise deleted. Imputed data will be biased because item missing is not likely to be random. Person groups who are typically underrepresented in surveys such as the less interested or socially excluded, or the less educated, are probably concerned by item non-

response to a higher extent. This is the case, in particular, when the questionnaire is difficult and/or long. Therefore, there should be measures to reduce item non-response to the greatest extent possible: From this research we deduce that questions should be easy to understand (they do not need to be short), should not require too much memory or, in web surveys, calculations that are too difficult, should have fewer response options in telephone surveys, and questionnaires should be short or asked in several parts, especially in telephone surveys. Special care should be taken to avoid item refusals in web surveys as much as possible. Still, the web mode seems to produce more DK answers and item refusals independent of respondent and question characteristics. Therefore, questionnaire designers should use all available measures to increase the chances of getting a valid answer in web surveys.

This study has a number of limitations. Our result is somewhat limited in the sense that the groups we compare are not exclusively web- or telephone-based interviews. As we have mentioned, respondents might have answered parts of the survey (the household questionnaire, proxy interviews for children below 14, etc.) using another mode. Further research could assign sample members to exactly one mode. In addition, it is likely that although we controlled for several socio-demographic variables, there may still be uncontrolled selection in the two modes. As a solution to this issue, one could assign respondents to two waves with a randomised order of the first and the second mode. Issues of selective initial non-response or attrition would need special attention. A more elaborated experiment such as that to investigate mode of interview effects was not feasible within the pilot study examined in this paper.

Our data did not allow for tests of effects of interactions of mode and respondent characteristics. It would be interesting to explore, for example, whether DK is higher for questions that are more difficult if respondents have a lower education. This may then show limits of complexity of

questions for certain person-groups. In addition, it would be interesting to improve our knowledge about the true meaning of a DK or a mid-scale response. Some of them may be a valid response, some may be a form of satisficing. In our research, it was impossible to decide which DK or mid-scale answer was based on satisficing or which was a true answer. To a limited degree, this is also true for item refusals. A better knowledge of the actual meaning would enable survey designers to produce adapted answer categories.

Overall, we think that our data represents a real-life setting very well and provides new insight concerning which types of questions work better or worse in web surveys when compared with more traditional survey modes.

References

Atkeson, L.R. and Adams, A.N. 2018. "Mixing Survey Modes and Its Implications," in: *The Oxford Handbook of Polling and Survey Methods*, edited by. Atkeson, L.R. and R.M. Alvarez. New York: Oxford University Press.

Bassili, J. N. 1996. "The How and the Why of Response Latency Measurement in Telephone Surveys," in: Answering Questions: Methodology for Determining Cognitive and Communicative Processes in Survey Research, edited by Schwarz, N. and S. Sudman, 319-346. San Francisco: Jossey-Bass.

Bowling, A. (2005). "Mode of questionnaire administration can have serious effects on data quality," *Journal of Public Health* 27: 281–291.

Cannell, C.F., Miller, P.V. & L. Oksenberg. 1981. "Research on Interviewing Techniques," in: *Sociological Methodology*, edited by S. Leinhardt, 389-437, San Francisco: Jossey-Bass.

Christian, L.M. and D.A. Dillman. 2004. "The Influence of Graphical and Symbolic Language Manipulations on Responses to Self-Administered Questions," *Public Opinion Quarterly*, 68(1): 57-80.

Couper, M.P. & F. Kreuter. 2013. "Using Paradata to Explore Item-Level Response Times in Surveys," *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 176(1): 271-286.

de Leeuw E. 1992. "Data quality in mail, telephone and face to face survey." Dissertation, university of Amsterdam, available at: http://edithl.home. xs4all.nl/pubs/disseddl.pdf

de Leeuw, E. 2005. "To mix or not to mix data collection modes in surveys," *Journal of Official Statistics*, 21(5): 233-255

Edwards, W.S. and D. Cantor. 2004. "Towards a Response Model in Esablishment Surveys," in: *Measurement Errors in Surveys*, edited by Biemer, P.P., Groves, R.M., Lyberg, L.E., Mathiowetz, N.A., and S. Sudman, 211-236, Hoboken, NJ: John Wiley & Sons.

Fielding, A. and Goldstein, H. 2006. "Cross-classified and multiple membership structures in multilevel models: An introduction and review," University of Birmingham.

Fowler, F.J., Jr., Roman, A.M., and Z.X. Di. 1998. "Mode Effects in a Survey of Medicare Prostate Surgery Patients," *Public Opinion Quarterly*, 62(1): 29-46.

Fricker, S., Galesic, M., Tourangeau, R., and T. Yan. 2005. "An Experimental Comparison of Web and Telephone Survey," *Public Opinion Quarterly*, 69, 370-392.

Goodman, A., Brown, M. Silverwood, R.J., Sakshaug, J.W., Calderwood, L., Williams, J. and Ploubidis, G.B. 2020. "The impact of using the Web in a mixed mode follow-up of a longitudinal birth cohort study: evidence from the National Child Development Study", *CLS Working Paper* 2020/9. London: UCL Centre for Longitudinal Studies.

Greene, J., Speizer, H., and W. Wiitala. 2008. "Telephone and web: Mixed-mode challenge," *Health services research*, 43(1): 230-248.

Groves, R.M., Fowler, F.J., Couper, M. P., Lepkowski, J. M., Singer, E., and Tourangeau, R. 2011. *Survey Methodology*. Hoboken: John Wiley and Sons.

Guzy, N., and Leitgöb, H. 2015. "Assessing mode effects in online and telephone victimization surveys," *International Review of Victimology*, 21(1): 101-131.

Harkness, J., Van de Vijver, F.J.R., and P. Ph. Mohler. 2003. *Cross-Cultural Survey Methods*. Hoboken, NJ: Wiley.

Holbrook, A., Cho, Y.I., and T. Johnson. 2006. "The Impact of Question and Respondent Characteristics on Comprehension and Mapping Difficulties," *Public Opinion Quarterly*, 70(4): 565-595.

Hope, S., Campanelli, P., Nicolaas, G., Lynn, P., and A. Jäckle. 2014. "The Role of the Interviewer in Producing Mode Effects: Results from a Mixed Modes Experiment Comparing Face-to-Face, Telephone and Web Administration", No. 2014-20, *ISER Working Paper Series*. (Available from https://www.iser.essex.ac.uk/research/publications/working-papers/iser/2014-20.pdf.)

Johnson, T.P., Cho, Y.I., Holbrook, A.L., O'Rourke, D.P., Warnecke, B., and N. Chavez. 2006. "Cultural Variability in the Effects of Question Design Features on Respondent Comprehension of Health Surveys," *Annals of Epidemiology*, 16: 661-668.

Kleiner, B., Lipps, O. & E. Ferrez. 2015. "Language Ability and Motivation Among Foreigners in Survey Responding," *Journal of Survey Statistics and Methodology*, 3(3): 339-360.

Krosnick, J.A. 1991. "Response Strategies for Coping with the Cognitive Demands of Attitude Measures in Surveys," *Applied Cognitive Psychology*, 5: 231-236.

Krosnick, J.A., Narayan, S., and W.R. Smith. 1996. "Satisficing in Surveys: Initial Evidence", *New Directions for Evaluation*, 70: 29-44.

Leigh, J. H., & Martin Jr, C. R. 1987. "Don't Know" Item Nonresponse in a Telephone Survey: Effects of Question Form and Respondent Characteristic, *Journal of Marketing Research*, 24(4): 418-424.

Lenzner, T., Kaczmirek, L, and A. Lenzner. 2009. "Cognitive Burden of Survey Questions and Response Time: A Psycholinguistic Experiment," *Applied Cognitive Psychology*, 24: 1003-1020.

Lesser, V., Newton, L., and Yang, D. (2012). "Comparing Item Nonresponse across Different Delivery Modes in General Population Surveys". *Survey Practice*, 5(2).

Narayan, S. and J.A. Krosnick. 1996. "Education Moderates Some Response Effects in Attitude Measurement," *Public Opinion Quarterly*, 60: 58-88.

Peytcheva, E. 2008. *Language of Administration as a Source of Measurement Error: Implications for Surveys of Immigrants and Cross-Cultural Survey Research.* Ann Arbor, MI: Pro-Quest, UMI Dissertation Publishing.

Revilla, M. (2012). "Impact of the mode of data collection on the quality of survey questions in social sciences". Doctoral dissertation, Universitat Pompeu Fabra, Barcelona.

Roberts, C., and Vandenplas, C. (2017). "Estimating components of mean squared error to evaluate the benefits of mixing data collection modes," *Journal of Official Statistics*, 33(2): 303-334.

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

Schuman, H. 1992. "Context Effects: State of the Past/State of the Art." In: *Context Effects in Social and Psychological Research*, edited by Schwarz, N. and S. Sudman, 5-20. New York: Springer-Verlag.

Schuman, H. and S. Presser. 1981. *Questions and Answers in Attitude Survey: Experiments on Question Form, Wording and Context*. New York Academic Press.

Simon, H.A. 1957. *Models of Man*. New York: Wiley.

Smyth, J.D., Dillman, D. Christian, L.M., and M.J. Stern. 2006. "Effects of Using Visual Design Principles to Group Response Options in Web Survey," *International Journal of Internet Science*, 1: 6-16.

Sturgis, P., Roberts, C., and Smith, P. 2014. Middle alternatives revisited: how the neither/nor response acts as a way of saying "I don't know," *Sociological Methods & Research*, 43(1): 15-38.

Sudman, S., Bradburn, N.M., and N. Schwarz. 1996. *Thinking about Answers*. San Francisco, CA: Josey-Bass.

Tillmann, R., Voorpostel, M., Antel, E., Kuhn, U., Lebert, F., Ryser, V.-A., Lipps, O. and Wernli, B. 2016. "The Swiss household panel study: Observing social change since 1999," *Longitudinal and life course studies*, 7(1): 64-78. doi:10.14301/llcs.v7i1.360

Tourangeau, R., Couper, M., and F. Conrad. 2004. "Spacing, Position and Order: Interpretive Heuristics for Visual Features of Survey Questions," *Public Opinion Quarterly*, 68(3): 368-393.

Tourangeau, R., Rips, L.J., and K. Rasinski. 2000. *The Psychology of Survey Response*. New York: Cambridge University Press.

Tourangeau, R., and T. Yan. 2007. "Sensitive Questions in Surveys," *Psychological Bulletin*, 133(5): 859-883.

Voorpostel, M., Kuhn, U., Tillmann, R., Monsch, G.-A., Antal, E., Ryser, V.-A., Lebert, F. and Klaas, H. 2020. Introducing web in a refreshment sample of the Swiss Household Panel: Main findings from a pilot study. Fors Working Series paper 2020-02. Lausanne: FORS.

Warnecke, R.B., Johnson, T.P., Chavez, N., Sudman, S., O'Rourke, D.P., Lacey, L., and J.Horm. 1997. "Improving Question Wording in Surveys of Culturally Diverse Populations," *Annual Epidemiological Psychology,* 7:334-342.

Yan, T. & R. Tourangeau. 2008. "Fast Times and Easy Questions: The Effects of Age, Experience, and Question Complexity on Web Survey Response Times," *Applied Cognitive Psychology,* 22(1): 51-68.

Young, R. 2012. Don't Know Responses in Survey Research. Unpublished dissertation. Department of Sociology: Pennsylvania State University.

Table 1. Effect of respondent and question characteristics on item non-response by mode

	Model 1 Don't knows telephone	Model 2 Don't knows web	Model 3 Refusal telephone	Model 4 Refusal web	Model 5 Mid-scale telephone	Model 6 Mid-scale web
	b	b	b	b	b	b
Question characteristics: Content						
Memory med.	ns	ns	ns	ns	2.77**	2.65**
Memory high	2.34**	1.00*	ns	ns	-	-
Calculation	ns	1.01**	ns	ns	ns	ns
Demographic	ns	ns	ns	ns	-	-
Factual	ns	ns	ns	ns	-	-
Question characteristics: Form						
Question length	-.41*	ns	ns	ns	ns	ns
2 responses	-1.62**	ns	-.99**	ns	-	-
3 responses	-2.18**	ns	-4.09*	.693*	-	-
Placement mid	1.08*	ns	1.10*	1.07**	ns	ns
Placement end	2.49**	1.38**	1.26**	.89**	ns	.62**
Respondent characteristics						
Survey language (Ref. French)						
Swiss German	-.52**	-.13**	ns	-.16**	ns	ns
Italian	ns	.22**	1.13**	.41**	ns	ns
Language (Ref. first best)						
Second best	.22*	ns	-.83**	ns	ns	ns
Neith. 1st nor 2nd	.37*	.19*	-1.65**	1.20**	ns	ns
Nationality (Ref. Swiss)						
Nationality 2	.44**	ns	.62*	ns	ns	ns
Nationality 3	ns	.592**	2.17**	.49**	ns	ns
Age (Ref. 18-29)						
30-39	ns	-.47**	ns	-.59**	ns	.17*
40-49	-.33*	-.70**	ns	-.91**	ns	.21**
50-59	-.30*	-.88**	ns	-.75**	ns	.23**
60-69	ns	-.92**	ns	ns	ns	.16**
70+	ns	-.87**	ns	ns	.242**	.19**
Education	-.17**	-.15**	.13**	-.15**	ns	-.02**
Male	-.37**	-.21**	ns	-.16**	ns	ns
Reference Person	ns	ns	.65**	ns	ns	ns
G2 (Tel>Web))	.35**	ns	.65**	-.18**	ns	ns
Constant	-5.49**	-4.14**	-8.12**	-3.51**	-2.55**	-2.63**
Observations	83'296	199'570	83'296	199'570	27'186	65'490
Bayesian DIC	9'052	40'851	3'196	50'141	20'063	47'936

NOTES: *p<0.05, ** p<0.01. Data: Swiss Household Panel SHP IV pilot sample wave 1, no f2f, only HH with telephone (those with no telephone are not in the telephone sample in group 1).