Article

Social Science Computer Review
2016, Vol. 34(1) 116-132

DOI: 10.1177/0894439315573744
ssc.sagepub.com

Handling Do-Not-Know Answers: Exploring New Approaches in Online and Mixed-Mode Surveys

Edith D. de Leeuw[1], Joop J. Hox[1], and Anja Boevé[2]

Abstract

An important decision in online and mixed-mode questionnaire design is if and how to include a "do-not-know" (DK) option. Mandatory response is often a default option, but methodologists have advised against this. Several solutions for the DK category are suggested. These include (1) not explicitly offering a DK, but skipping questions is allowed, (2) explicitly offering a DK option with visual separation from the substantive responses, and (3) using the interactivity of the web to emulate interviewer probing after a DK answer. To test these solutions, experimental data were collected in a probability based online panel. Not offering DK, but allowing respondents to skip questions, followed by a polite probe when skips occurred, resulted in the lowest amount of missing information. To assess the effect of probing across different modes, a second experiment was carried out that compared explicitly and implicitly offering the DK option for web and telephone surveys.

Keywords

data quality, emulating interviewers, item nonresponse, probing, visual design, web survey

Missing data or item nonresponse is considered an important indicator of data quality in surveys (Groves, 1989). Item nonresponse occurs when a respondent (e.g., panel member) partially provides data and thus does not fully complete the survey. Consequently, for particular questions, no data are available and the data matrix displays gaps, which are usually indicated by a missing data code. Both survey methodologists and survey statisticians have studied causes of and solutions for item nonresponse, and several design choices that may influence item nonresponse (e.g., survey mode, question design, question difficulty and respondent characteristics; De Leeuw, Hox, & Huisman, 2003; Groves, Dillman, Eltinge, & Little, 2002).

Survey mode and implementation influence the amount of item nonresponse. A well-trained interviewer, for instance, can follow up a "do-not-know" (DK) response with a polite probing question to encourage a meaningful response, while in a paper-based mail survey the respondent is fully

[1] Utrecht University, Utrecht, the Netherlands
[2] University of Groningen, Groningen, the Netherlands

Corresponding Author:
Edith de Leeuw, Plantage Doklaan 40, NL 1018 CN, Amsterdam, the Netherlands.
Email: e.d.deleeuw@uu.nl

in control and may decide to skip a question. An often cited advantage of computer-assisted and online surveys is that responses can be automatically checked, and respondents can be required to answer a survey question before they can continue. However, making responses mandatory may also cause more breakoffs or unwanted response behavior such as satisficing (Tourangeau, Conrad, & Couper, 2013).

Empirical mode comparisons support the idea that interviewer probing is important in reducing missing data. In an early meta-analysis, De Leeuw (1992) for instance, found that self-administered mail surveys produced a higher amount of item missing data than comparable face-to-face and telephone interviews. For comparisons with web surveys, the results are less clear cut, depending on the design of the surveys. Web surveys, like mail surveys, where skipping a question or explicitly answering DK is allowed, result in more item missing data than comparable interviews where a nonsubstantive answer is probed (e.g., Heerwegh & Loosveldt, 2008). However, if the design of a face-to-face interview allows a DK answer to be directly accepted by interviewers without any probing, the opposite may occur and web surveys result in less item nonresponse (e.g., Fricker, Galesic, Tourangeau, & Yan, 2005). For a summary of this literature, see De Leeuw and Hox (2011, pp. 59–61). The design and implementation of DK is therefore especially important in mixed-mode surveys, where web and interviews are combined (De Leeuw, 2005). For instance, Jaeckle, Lynn, and Burton (2013) found in an experiment that a mixed-mode design combining web and face-to-face interviews produced more item nonresponse than a single mode face-to-face interview, which was entirely due to the larger amount of missing data in the web part of the mixed-mode survey. Reducing missing data in web surveys is very important for the design of mixed-mode web interview surveys, where it is desirable to achieve similar data quality in each mode (Al Baghal & Lynn, 2014).

Although explicitly including a DK option may result in increased amounts of missing data, survey designers face several other important considerations concerning the decision whether or not to explicitly include a DK option. Depending on the question topic and difficulty, and on respondent characteristics, respondents may genuinely not know the answer, and not giving them the opportunity to express this may threaten the reliability and validity of the data (Converse, 1976; Converse & Presser, 1986). On the other hand, explicitly offering a DK option may give respondents an easy way out and invite satisficing (Krosnick et al, 2002). It is therefore common in interviews not to offer a DK option explicitly. Interviewers are moreover instructed to allow DK answers, but to also follow up these nonsubstantive DK answers with a polite probe (Sudman & Bradburn, 1982). This is also often the standard option in software for computer-assisted telephone and face-to-face interviews (Bethlehem & Biffignandi, 2012).

Several measures have been proposed to reduce the number of DK answers in online surveys. Making a response mandatory automatically reduces the amount of item missing answers to zero, but may have other undesirable effects. Both Couper (2008) and Dillman, Smyth, and Christian (2009) argue that a mandatory response goes against the ethical norm of voluntariness in a survey and may annoy respondents. In order to minimize respondents' irritation, Couper (2008) argues that error messages implemented in an online survey should be polite and helpful. Therefore, when designing a web survey with mandatory responses, one should take care to implement a "polite error message" or soft error prompt.

When responses are not mandatory in an online survey, the decision remains whether or not to explicitly offer a DK answer. In both cases, the visual screen display is an important consideration. Tourangeau, Couper, and Conrad (2004) advise visually separating substantive scale points from nonsubstantive response options because otherwise the visual midpoint of the response scale is altered, which changes the response distribution. Visually separating the DK option from the substantive response options may also reduce the number of DK answers (Dillman, Smyth, & Christian, 2009), although the empirical evidence is scarce (DeRouvray & Couper, 2002; Tourangeau et al., 2013).

Another potential solution is to implement a polite error message, rather like an interviewer prompt. For instance, in a web survey, DeRouvray and Couper (2002) prompted respondents who *skipped* a question by asking them to select the <back button> and answer the question. This did reduce the number of skip attempts and thus the average missing value rate. However, the use of an explicit decline-to-answer option was not reduced because this particular option was *not* followed by a prompt (DeRouvray & Couper, 2002, p.6). A more sophisticated prompt was used by Wine, Cominole, Heuer, and Riccobono (2006) who went a step further and emulated an interviewer probe after a nonsubstantive answer to reduce item missing data. Wine et al. (2006) had to change their main mode from telephone to web surveys and used existing interviewer texts from previous telephone studies to write a polite, system-generated message to motivate and probe the respondent for more information.

Several studies used the interactive potential of web surveys to provide respondents with customized feedback to motivate them in a similar way interviewers do. For instance, Smyth, Dillman, Christian, and McBride (2009) examined the effect of different visual designs of answer boxes and motivational introductions to responses for open-ended questions, showing that open questions can be an effective format in web surveys as they elicit descriptive information from respondents. In the context of open-ended questions, Oudejans and Christian (2011) found that adding a motivational statement about the importance of a question and follow-up probes helped to reduce item nonresponse and improved the richness of information provided. Even in a qualitative online interview, probes can be implemented successfully, as Behr, Kaczmirek, Bandilla, and Braun (2012) showed in a study on probing in online cognitive interviews.

Experimental research on closed-ended questions is scarce. DeRouvray and Couper (2002) found that a probe after a respondent skipped a closed question reduced the overall amount of item nonresponse. Al Baghal and Lynn (2014) investigated two types of motivational probes, namely, a reactive probe directly after a question was left unanswered and a probe placed at the end of the questionnaire asking respondents to answer a couple of important questions that were previously left unanswered. The direct, reactive probe was far more effective than the postponed probe and reduced item nonresponse in web surveys to the level of face-to-face interviews.

In this study, we experimentally investigated which approaches work best in reducing missing data in an online survey, an implicit (i.e., not offered, but skipping allowed) DK option, an explicit DK option, using two different approaches to visually separate it from the substantive answers, and emulating a polite interviewer probe after a DK answer. We also investigated whether DK followed by an online interviewer-emulated probe works as well as a real interviewer probe, which is extremely important for designing mixed-mode web interview surveys (Al Baghal & Lynn, 2014).

The objective of our study is to find the best way to handle DK answers in online surveys. Two criteria are used to define "best": first, the amount of item nonresponse, and second, the reliability of the answers given. We also investigated if probing influenced respondents' evaluation of the survey. Our research questions and hypotheses are:

1. Which online question format produces the least item nonresponse, and what is the influence on the reliability of the data and respondents' survey evaluation? We hypothesize that explicitly offering DK will increase the amount of item nonresponse. We have no explicit hypotheses on the effects which eliciting DK responses will have on scale reliability and respondents' survey evaluation.
2. If one has to offer DK, what is the best way to do this, and which of the options offered is best, regarding reliability of responses and respondents' evaluation? We hypothesize that visual layouts in the web version that separate the DK option on the screen from the substantive responses work better for all criteria.

3. Does emulating a polite interviewer probe reduce nonsubstantive answers, and how does it influence the reliability of the answers and respondents' evaluation? We hypothesize that probing will reduce the amount of item nonresponse. As Al Baghal and Lynn (2014) found hardly any effect of probing on the survey estimates, but Oudejans and Christians (2011) found a slightly negative effect, we have no explicit directional hypotheses on the effects of probing on scale reliability and respondents' evaluation of the questionnaire.
4. What are the consequences for mixed-mode surveys, and more specifically, does an emulated polite interviewer probe on the web work the same as in a telephone interview? We hypothesize that explicitly offering DK has a stronger effect on the amount of DK responses in the web mode than in the telephone mode and that an emulated probe is likely to be less effective than a probe by an interviewer. We expect that in the web mode, probing does not reduce the amount of DK responses to the level that is reached when DK is not offered as an explicit option.

Method

Data

The data were collected among members of the LISS panel, a probability-based online panel of the Dutch population. The LISS panel was established in 2007 and is based on a probability sample of households drawn from the population register by Statistics Netherlands. Households that could not participate otherwise were provided with a computer and broadband connection (Scherpenzeel & Das, 2011). Each month panel members receive one or more questionnaires with an average monthly burden of between 30 and 40 min. Panel members are usually sent two reminders and receive a monetary incentive for each questionnaire they respond to. More information about the LISS panel can be found at www.lissdata.nl.

A random sample of 6,134 panel members received a questionnaire online as usual, while a different random sample of 2,000 panel members were approached by telephone by a professional fieldwork organization that also does the recruiting interviews for the LISS panel. Online, 4,003 (65%) panel members responded and over the telephone 1,207 (60%).

As the response rates between the two modes differed, we checked if differential nonresponse compromised the comparability between the modes. For all panel members, biographical and psychographic information was available, and small, but significant, differences between modes were found for the variables age and household size. Propensity score weights were constructed and all analyses comparing telephone and web respondents were carried out on both weighted and unweighted data. Since results did not differ, only the unweighted results will be reported here (see De Leeuw, Hox, & Scherpenzeel, 2010).

Questionnaire

The topic of the questionnaire was use and acceptability of advanced medical technology. The questionnaire contained six questions on the use of embryos in medical research, which in previous self-administered surveys showed a high percentage of item nonresponse (Steegers, Dijstelbloem, & Brom, 2008). Although the format of the questions is designed to ask factual information, the actual information asked for is generally not expected to be available to the respondents. Thus, these six questions in fact ask for opinions on the state of medical technology about which most respondents are expected to experience considerable uncertainty. For example, do you think that embryos are used in scientific medical studies searching for cures for severe diseases? Response categories were "yes" and "no" and depending on the experimental manipulation a DK option was offered (for all six questions see Appendix A1).

Design of the Experiments

Online survey. In the first experiment, we investigated different ways of handling the DK option in an online survey. A two by three experimental design was used. The first factor manipulated accepting a DK directly versus only accepting it after a polite probe (see Appendix A2). The second factor manipulated no explicit offering versus offering DK in two different ways, either visually separating DK by an extra empty line space or offering DK as a special button. Examples of screenshots are presented in Appendix A3. This led to six experimental groups. In all six groups, respondents could proceed to the next question without giving a substantive answer (i.e., skip). A seventh experimental group, the mandatory answer group, was added as baseline comparison group. In this group, the respondents had to answer the question with a substantive answer before they were allowed to proceed to the next question. The standard LISS error message was used which repeats the question with the prompt: "You have not given an answer" (see Appendix A3 for a screenshot). Whether or not responses are mandatory in LISS questionnaires is left to the decision of the substantive researchers and is not a panel policy. In most LISS studies, however, providing a response is mandatory. Online respondents, drawn from the LISS panel, were randomly assigned to the seven experimental conditions.

Comparison with telephone survey. In the second experiment, we assess the effect of probing for a mixed-mode design. We investigated if a polite probe after a DK answer in web surveys is as effective as a probe in a telephone interview. A two by two design was used. The first factor indicates survey mode: web survey versus telephone interview. The second factor manipulates not offering DK versus explicitly offering DK. A DK response was in this experiment *always* followed by the polite probe (Appendix A2). Respondents were drawn from the LISS panel and randomly assigned to the different conditions. In the telephone condition, randomization was done within interviewers. Trained interviewers were used who received a brief additional training for this study and were monitored during data collection.

Analysis and Results

We investigated the influence offering DK had on three different quality aspects, namely, the amount of item nonresponse, the reliability of the answers, and the overall survey evaluation by the respondents. For the latter, the standard evaluation questions of the LISS panel were used. At the end of each questionnaire, respondents evaluate the questionnaire in terms of difficulty and clarity and the extent to which it was interesting, pleasant, and thought provoking. For these questions, the 5-point answer scale ranges from "clearly yes" to "clearly no."

DK options and Probing in Online Survey

Item nonresponse. To analyze the effect of the different experimental conditions on the amount of item nonresponse, a general linear model with a Poisson distribution and log link was applied, using maximum likelihood estimation and robust confidence interval estimation. The dependent variable was a count variable, that is, the sum of missing values over the 6 items. Independent variables were the six experimental conditions, coded as five dummy variables for all contrasts and interactions. The six experimental conditions were (1) no explicit DK offered, but skipping allowed, no probe; (2) no explicit DK offered, but skipping allowed, followed by polite probe; (3) explicit DK offered, visually separated by extra space, no probe; (4) explicit DK offered, visually separated by extra space, polite probe; (5) DK offered as special button, no probe; and (6) DK offered as special button, polite probe. The seventh mandatory (required) answer group was excluded from this analysis, as the

Table 1. For Each Experimental Condition: Number of Respondents (*N*), Percentage of Total Missings (Item Nonresponse) Across the Six Questions, Reliability of 6-Item Scale (α) With Listwise Deletion (LD) and Based on EM Estimated Covariances.

Experimental Condition	*N*	% Total Item Nonresponse	Reliability (α) LD	Reliability (α) EM
1. No explicit DK (skipping allowed)	617	1.0	.65	.68
2. No explicit DK + Probe	574	0.5	.74	.75
3. DK visually separated by extra (empty line) space	544	33.3	.51	.52
4. DK empty line + Probe	569	8.8	.67	.67
5. DK special button	587	24.4	.66	.65
6. DK button + Probe	549	8.1	.64	.67
7. Mandatory response	563	0% per definition	.71	.71

Note. DK = do-not-know. If the condition included a probe, the reported statistics are reported on the results after probing.

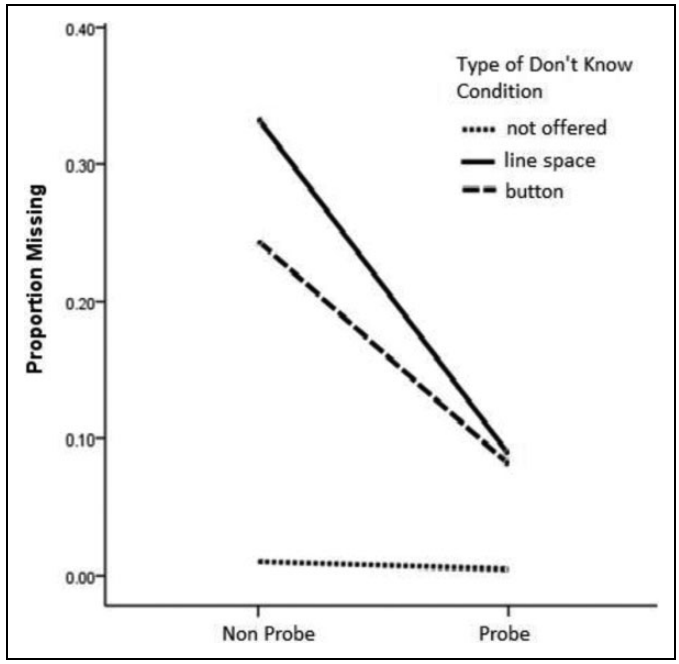


Figure 1. The proportion of item nonresponse across 6 items for six experimental conditions (*after* probing). The Mandatory Response Condition was excluded from the analysis, since, by definition, it does not contain item nonresponse.

amount of missing information for this baseline group is by definition zero. Table 1 presents the mean percentages of DK answers per group after probing. The results are graphically summarized in Figure 1.

We found clear main effects of offering DK ($p < .001$) and of probing ($p < 001$). Offering DK significantly increases the fraction of DK responses, using a probe decreases this fraction, but not enough to completely negate the effect of explicitly offering DK. How the DK option was visually presented on the screen had both a significant main effect ($p < .001$) and an interaction effect with using a probe ($p = .02$). Figure 1 makes clear what happens: The presentation (extra line space

before the option versus a separate button) makes a difference only if no probing is used (using a button leads to less DK responses than having an empty line), after probing there is no difference between the two.

When we directly interpret the amount of DK answers presented in Table 1, we conclude that not offering a DK option but allowing respondents to skip the question leads to the least amount of missing data (1.0% without probe, 0.5% with a probe). Using the DK as a separate button leads to less missing information (24.4% without probe and 8.1% with a probe) than separating the DK by an empty line (33.3% without a probe and 8.8% with probe). A polite interviewer-emulated probe significantly reduced the final amount of missing data in all conditions. Not explicitly offering DK with skipping allowed produced fewer missing data even without probing than explicitly offering DK with probing (1% vs. 8–9%). Not offering a DK followed by a polite probe resulted in the lowest amount of missing information (0.5%).

Reliability. The way DK options are offered in a web survey and whether or not DK answers are probed do reduce the amount of missing data. However, it is also important to ensure that data quality does not suffer from efforts to reduce missing information and using follow-up probes. Hence, the influence of the way DK was offered and probing versus nonprobing on the reliability was investigated and compared to the reliability of the answers in the mandatory answer condition. For all seven experimental conditions, we calculated the reliability coefficient α over the six questions used in this experiment, where applicable after probing. Coefficient α indicates the consistency with which the six questions are answered, and here it reflects a general perception that embryonic material is or is not used for a variety of research purposes. A high coefficient α implies absence of random measurement error but not necessarily a high validity. As stated in the introduction, we expect that not giving respondents the opportunity to choose DK will lower reliability. Nevertheless, the comparison of different reliability coefficients must be regarded as exploratory.

Standard statistical programs use listwise deletion when calculating reliability coefficient α, meaning that respondents with a missing value on one or more questions are completely excluded from the calculations. Since our experimental manipulations were designed to affect the probability of producing DK answers, which are coded as missing values for the reliability analysis, comparing the reliability under different conditions using listwise deletion could cause a considerable bias in the results. We therefore also estimated the scale reliability including the partially available information from subjects with incomplete data, by calculating coefficient α on the interitem covariances estimated using the Expectation Maximization (EM) estimation method, available in SPSS Missing Value Analysis. This allowed us to calculate coefficient α in such a way that we did not need to remove a subject from the calculations because one or more of the 6 items was missing, which is the case in standard coefficient α calculations with listwise deletion. Both the reliabilities estimated under listwise deletion and under EM are reported in Table 1; the differences between them are very small and follow the same pattern across experimental conditions. The EM estimates are theoretically superior because they are based on more information and make weaker assumptions, so these were used (after a normalizing Z-transformation) as dependent variables in a fixed effect meta-regression (Hox, 2010, p. 209) to assess whether the different manipulations affected the reliability of the scale formed by the 6 items. Table 2 presents the results of the meta-regression, omitting the (nonsignificant) interaction effects.

Contrary to expectations, explicitly offering a DK significantly reduced the reliability. Making responses mandatory and using probing both increases reliability. How the probe was implemented has no effect, and there were no significant interactions (smallest $p > .07$, interactions not included in the table).

Evaluation. At the end of the questionnaire, five standard questions were used to evaluate respondents' experience of the entire questionnaire. These items are Did you find the survey difficult? Did

Table 2. Meta-Regression Results Predicting (Z-Transformed) Coefficient α.

Predictors	Regression Coefficient (*SE*)	*p* Value	β
Mandatory response	.25 (.08)	.00	.64
Probing (vs. not probing)	.21 (.06)	.00	.71
DK offered (vs. not offered)	−.32 (.08)	.00	−.74
DK button (vs. line)	.11 (.08)	.17	.26

Note. DK = do-not-know; *SE* = standard error.

Table 3. Regression Results for the Evaluation Items: Regression Coefficients and (*SE*).

Predictors/Evaluation Question	Difficult	Made Think	Clarity	Interesting	Enjoyable
Mandatory response	.06 (.06)	.10 (.05)	.01 (.05)	.06 (.05)	.06 (.05)
Probing (vs. not probing)	.02 (.05)	.02 (.04)	−.01 (.04)	−.01 (.03)	.02 (.03)
DK offered (vs. not offered)	−.20* (.09)	−.21** (.07)	−.03 (.07)	−.13 (.07)	−.06 (.07)
DK button (vs. line)	.02 (.08)	−.04 (.07)	−.01 (.06)	−.01 (.06)	.02 (.06)
Probing × DK offered	.38** (.13)	.36** (.11)	.14 (.10)	.17 (.09)	.10 (.09)
Probing × Button	−.07 (.11)	.06 (.09)	.08 (.09)	.03 (.08)	.00 (.08)

Note. DK = do-not-know; *SE* = standard error.
*$p < .05$. **$p < .01$.

you find the topics of the survey interesting? Did the survey items make you think? Did you find the questions enjoyable? Did you find the questions clear? Each question was followed by a 5-point scale indicating agreement with the question. The two questions "did you find the questions difficult" and "did the questions make you think" were answered differently for the experimental conditions, while the other questions relating to enjoyment, clarity, or interest did not differ between the experimental conditions. Table 3 presents the regression coefficients for predicting the response to all five questions by the experimental conditions. The results show that when a DK was offered explicitly, the questions were experienced as less difficult to answer and less likely to make one think. However, for the condition where an explicit DK option was followed by a probe, respondents indicated that the questions were more difficult and made them think more. These results suggest that offering a DK without probing gives respondents an easy escape (less difficult, less thought provoking), while probing stimulates the question–answer process. The patterns of scale reliabilities presented earlier support this interpretation.

Comparing DK and Probing Online Versus Telephone

In the first experiment, we showed that in a web survey a polite interviewer-emulated probe is successful in reducing the amount of missing data without compromising the reliability of the data. For mixed-mode surveys, where the goal is to achieve similar data quality in each mode, it is important to know whether an online interviewer-emulated probe works just as well as a real probe in an interview survey. For this reason, we investigated the effect of probing after a DK answer both online Computer Assisted Web interview (CAWI) and via a Computer Assisted Telephone Interview (CATI). In both conditions, a split-ballot was conducted, in which a random half was not offered a DK option and another random half was explicitly offered a DK option. This led to four experimental groups (i.e., CATI + DK-offered, Online + DK, CATI no DK-offered, Online + No DK). All DK responses were followed up by a probe. The same questions and the same probes were used as in the first experiment.

Table 4. Number of DK Answers Before and After Probing, in Four Experimental Conditions.

	Before Probing	After Probing
CATI + DK-offered	0.91	0.28
Online + DK-offered	1.32	0.51
CATI + No DK	0.29	0.12
Online + No DK	0.05	0.02
Significance tests	Mode $p = .07$; DK $p < .00$; interaction $p < .00$	Mode $p = .05$; DK $p < .00$; interaction $p < .00$

Note. DK = do-not-know.

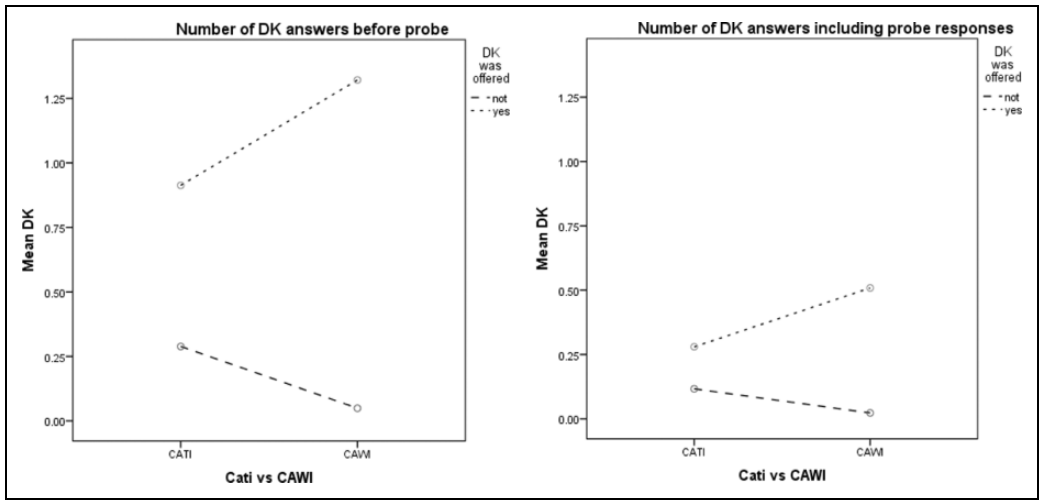


Figure 2. Mean number of do-not-know answers across 6 items before and after probing in telephone and online survey.

Item nonresponse. The number of DK responses over the six questions was calculated before and after probing. Table 4 presents the average number of DK answers in the four conditions, before and after probing.

Explicitly offering a DK option significantly ($p < .0001$) increases the amount of missing data before probing in both the telephone and the online survey. The effect of including a DK option clearly differs in telephone and online surveys (interaction $p < .0001$). When no explicit DK category was offered, respondents in the telephone condition offered more DK answers before probing than respondents in the online condition, but when an explicit DK option was offered this pattern reversed.

Using a polite probe after each DK response reduced the final amount of DK answers in both the telephone and online survey and repaired much of the damage done by including an explicit DK option, especially in the online survey (interaction $p < .0001$). To aid in the interpretation of the strong interaction effects, the number of DK answers in the two DK (not offered vs. offered) conditions and each mode (CATI vs. CAWI) are depicted in Figure 2 before and after probing.

Reliability. A polite probe does help to reduce the amount of DK answers in both telephone and online surveys, but are the probed answers as reliable as the spontaneous answers? As in the previous experiment, we investigated whether probing results in more guessing and random error and

Table 5. Reliability Coefficient α (EM) Before and After Probing in Four Experimental Conditions.

	Before Probing	After Probing
CATI + DK offered	.50	.50
Online + DK offered	.63	.65
CATI + No DK	.43	.44
Online + No DK	.74	.75
Significance tests:	Mode $p < .00$; DK $p < .00$; interaction $p < .03$	Mode $p < .00$; DK $p < .00$; interaction $p < .03$

Note. DK = do-not-know.

calculated the reliability coefficient α, using the EM estimation method, over the six questions for each mode before and after probing. The reliability coefficients and p values are shown in Table 5.

Overall, the reliability coefficients were not significantly different when comparing before and after probing ($p = .52$). However, we found an interesting mode effect ($p < .001$): Overall, the reliability in the online survey was higher than in the telephone survey, indicating that the answers given online were more consistent and contained less random error than over the phone. Offering an explicit DK has an interaction effect with mode: In the telephone mode, it increases the reliability and in the online mode it decreases the reliability, both before and after probing (both p values $< .01$).

Conclusion and Implications

This study adds an empirical basis to the debate on whether or not to offer a DK option in online and mixed-mode surveys. We show that explicitly offering a DK option in a web survey is not advisable. Offering a DK option leads to a considerable increase in the amount of missing data and often to a decrease in the reliability of the resulting answers. Not explicitly offering a DK option, but allowing respondents to skip a question followed by a polite probe works best: It leads to the lowest amount of missing data and the highest reliability.

When researchers find it necessary to explicitly offer a DK option, for theoretical or practical reasons, it is important to visually separate these clearly from the substantive answers. A special DK button which is more outside the foveal vision of the respondent works better by producing less missing data and a higher reliability than aligning the response categories with an extra empty space separating DK from the substantive answers. Using a separate button and following DK answers up with a probe leads to less missing information and higher reliability.

Consistent with these results, respondent evaluations suggest that offering a DK without probing gives an easy escape, since the questions are judged as less difficult and less thought provoking, while probing stimulates a more thorough question–answer process.

Surprisingly, the mandatory response condition, where respondents are forced to give an answer before they can proceed to the next question, also worked well and was comparable to the condition skipping allowed (no explicit DK) followed by a probe (% missing: 0 vs. 0.5%; reliability .71 vs. .75). It seems that the LISS error message in the mandatory condition also works as probe. The two conditions are much alike but show one big difference: the follow-up question (see Appendix A3). The initial screen is the same, but in the mandatory response condition the follow-up screen repeats the question with an error message saying "you have not given an answer," not allowing to proceed to the next question. In the skipping allowed condition, the follow-up screen provided a polite probe inviting the respondent to give a substantive answer but also allowing a DK if one really did not know the answer (see Appendix A3).

In the research literature, survey methodologists (e.g., DeRouvray & Couper, 2002) warn against a mandatory or forced response, as this is likely to increase break-offs from those who cannot or do not want to answer a specific question. We could not investigate this as our data set did not contain any break-offs, which may be the result of the reward that is offered to LISS-panel members for completing questionnaires. However, the LISS data include paradata on attempts to skip questions.

When we compared respondents in the mandatory condition who attempted to skip to respondents who directly gave an answer, we observed that respondents who attempted to skip items evaluated the difficulty in the questionnaire on average as 3.0 on a 5-point scale, while respondents who immediately gave an informative response evaluated the difficulty in the questionnaire as 2.5 ($p = .004$). Note that the evaluation questions refer to the entire questionnaire, not specifically to the questions that were part of the experiment, so the effect of making questions mandatory on the experienced difficulty of the questions is probably underestimated. The difference in evaluated difficulty level indicates that respondents who attempt to skip experience difficulty in the question–answer process and it may be better to give them the opportunity to indicate that, after more profound reflection they really do not know the answer, as is the case in the skipping allowed plus polite probe condition.

Finally, requiring a mandatory response to each question may not be ethical and may be contradictory to the requirements of Internal Review Boards that all individual answers to survey questions be "voluntary" (Dillman, 2012; Tourangeau et al., 2013).

Explicitly offering a DK option increases the amount of item nonresponse in both online and telephone surveys. Following DK answers up with a polite probe works well in reducing item nonresponse and is effective in both telephone and online surveys. Furthermore, probing did not negatively influence the reliability of the final answers. This means that emulating interviewers by probing in online surveys is not only possible but also successful.

One could argue that the LISS panel respondents are trained panel respondents which may compromise the generalizability of the results. The LISS panel is an infrastructure and the handling of question skipping is left to the researchers that use the panel. These in majority opt for making questions mandatory, which explains the very low level of question skipping when this is possible and DK is not explicitly offered. However, the respondents in the LISS panel are used to receiving a variety of question types on many different topics, and it is clear that when skipping questions is possible, they do use skipping. The same experiments presented here were replicated on a smaller scale in the newly established German PPSM-panels, where respondent training was not an issue, and the results presented here were replicated (de Leeuw & Hox, 2014).

The generalizability of our results is also limited due to the short list of questions used. In a list of questions for which skipping is allowed, but is followed up by a probe, the respondents who attempt to skip will learn that a skip is always followed by a probe. This may reduce the attempted skips later in a long list of questions. In our short list of six questions we could not analyze a learning effect. However, in longer lists there is a clear possibility that respondents who want to skip a question resort to satisficing, for instance by always agreeing, instead of attempting to skip and encounter a probe again.

The finding that the reliability tends to increase after probing does not rule out the possibility of satisficing. A higher reliability implies less random error, which is a nice result, but it does not imply a higher validity. The higher reliability could also be the result of systematic bias due to the use of response styles, such as acquiescence. We cannot rule this out; to distinguish between reliability and validity we need a different design, either by using a perfectly balanced scale (half positively and half negatively worded items, e.g., see Billiet & McClendon, 2000) or by using a multitrait-multimethod design (e.g., see Revilla & Saris, 2013).

One of the reviewers suggested that in the CATI condition interviewers might have an additional effect, for instance, because they differ in the degree to which they follow the protocol. Following up on this suggestion, we have done some additional multilevel analyses investigating interviewer

effects. We found that there were no significant interviewer effects on DK responses before probing. In the multilevel analyses, explicitly offering DK had a strong effect on producing DK responses before probing, consistent with the results presented earlier. The number of DK answers after probing shows the opposite effect: Interviewer effects are significant with an intraclass correlation of .09, and the effect of explicitly offering the DK option is no longer significant. This corroborates the results presented earlier and accentuates the importance of the interviewers in obtaining answers. Since we have no interviewer-level variables we cannot further explore interviewer effects. Nevertheless, the analysis of the interviewer effects in the CATI condition suggests that if DK answers are expected, it may be worthwhile to train interviewers in probing strategies, analogous to training interviewers for refusal conversion (cf. Groves & McGonagle, 2001). In our study, well-trained interviewers were used, who received a brief additional training for this specific study.

The results of these experiments clearly demonstrate that if multiple modes are employed in a survey, it is important to pay careful attention to question format and implementation details. In a mixed-mode design, one should optimize toward the mix, by using equivalent questionnaires and implementation procedures, in order to prevent avoidable mode effects. Based on the experiments of this study, we would *not* advise offering a DK option explicitly, but always allowing the respondents to skip a question online, just as they are allowed to spontaneously answer DK to an interviewer. In addition, a skip or spontaneous DK answer should always be followed up with a polite probe both over the telephone and online. Online, the interviewer is replaced by a polite system probe and one should take care that the same wording is used in both the interviewer delivered and online probe in order to make the final results for both modes as equivalent as possible.

Appendix

Topic of Questions and Text for Probe

A1: Questions used in the experiment (translation): for original Dutch questionnaire see Steegers, Dijstelbloem & Brom (2008) http://www.rathenau.nl/uploads/tx_tferathenau/Meer_dan_status_alleen_april_2008.pdf

"Do you think embryos are used for research on/into . . . "

- fighting off serious diseases
- development of reproductive (fertility) techniques such as IVF
- stem cells
- development of medical cures
- hereditary conditions
- development of embryos

response: yes/ no/ (DK)

A2: Text of probe for yes/no question used both online and over the telephone:

Thank you, we have recorded your answer. Perhaps you can express a preference for either yes or no? That would help us very much.

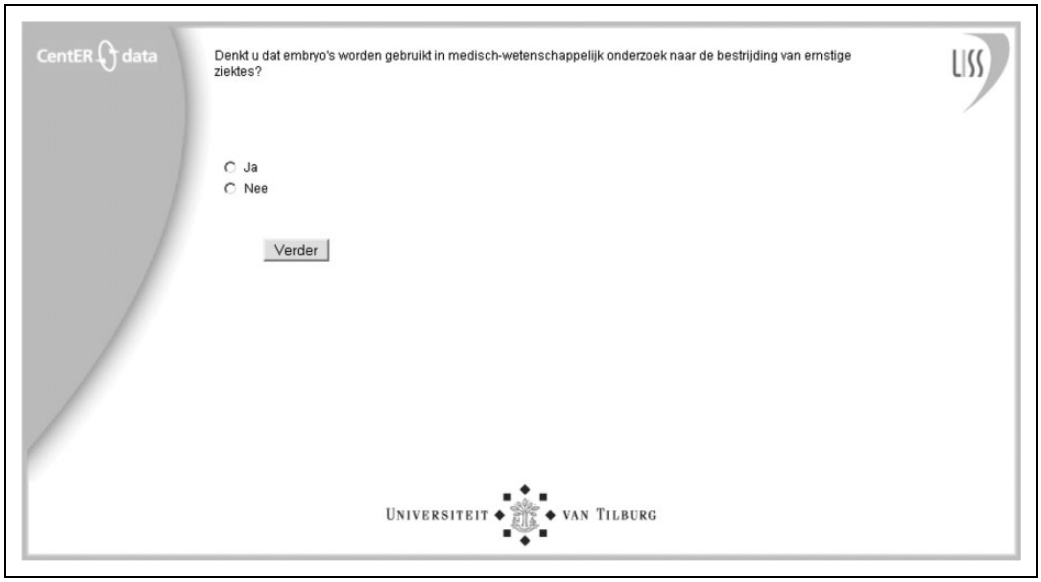
The question was <repeat question stem plus>

'I am inclined to say yes',

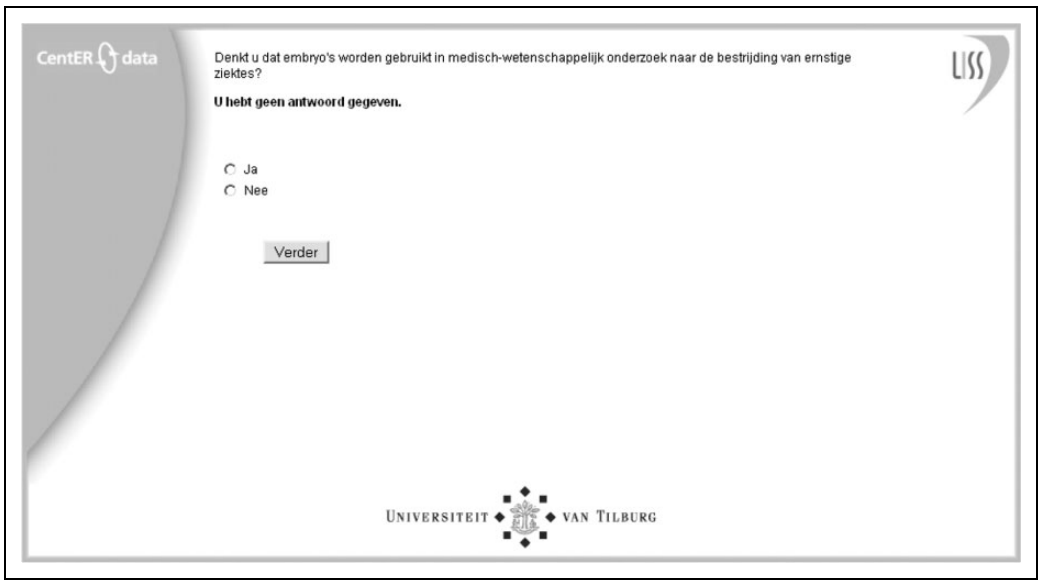
'I am inclined to say no',

'I really do not know'

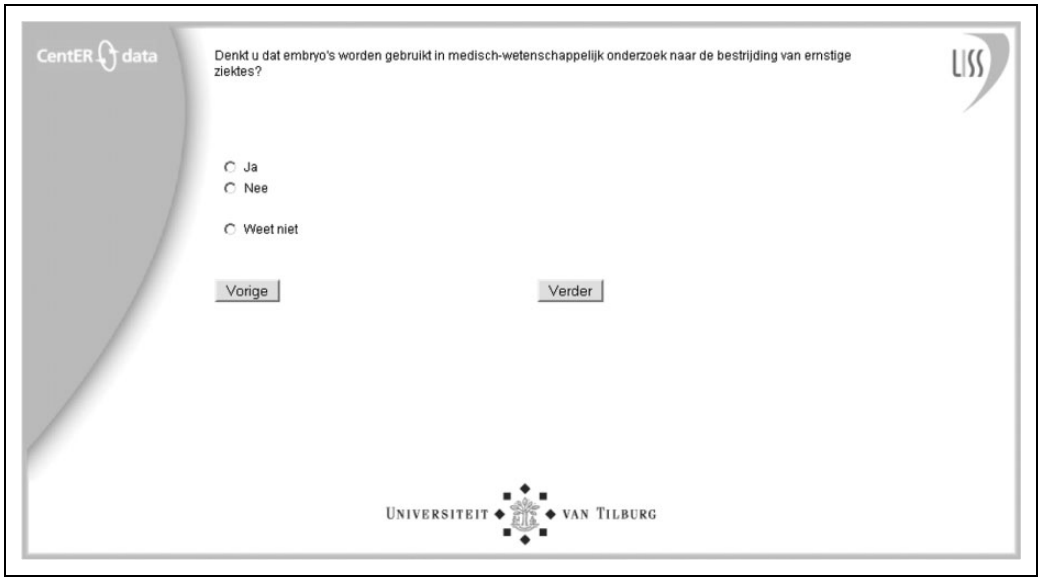
A3: Screenshots of experimental manipulations



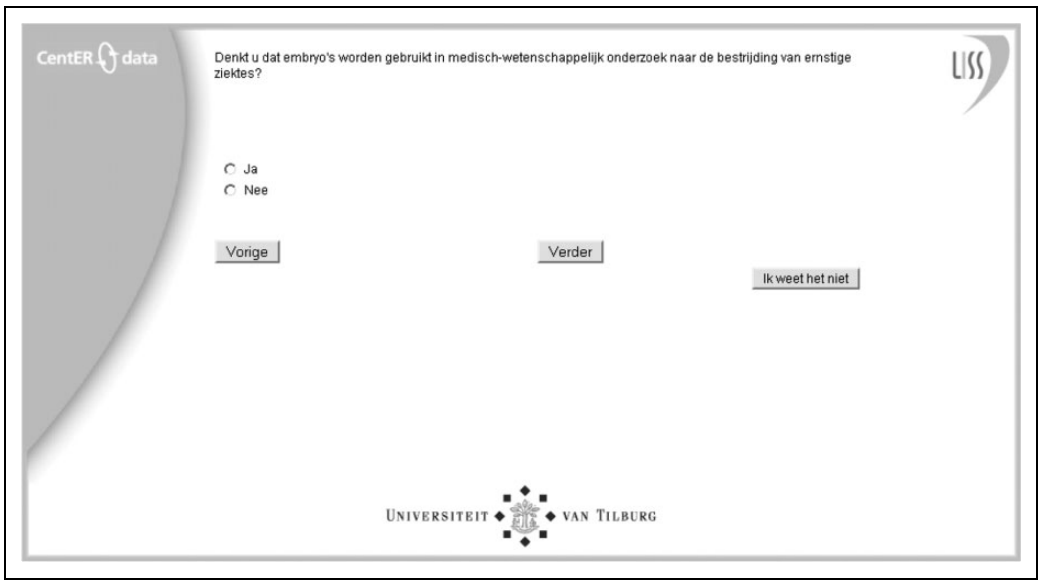
Example question: don't know not explicitly offered (responses Yes/No)
This format was used in two conditions: skipping allowed and mandatory response



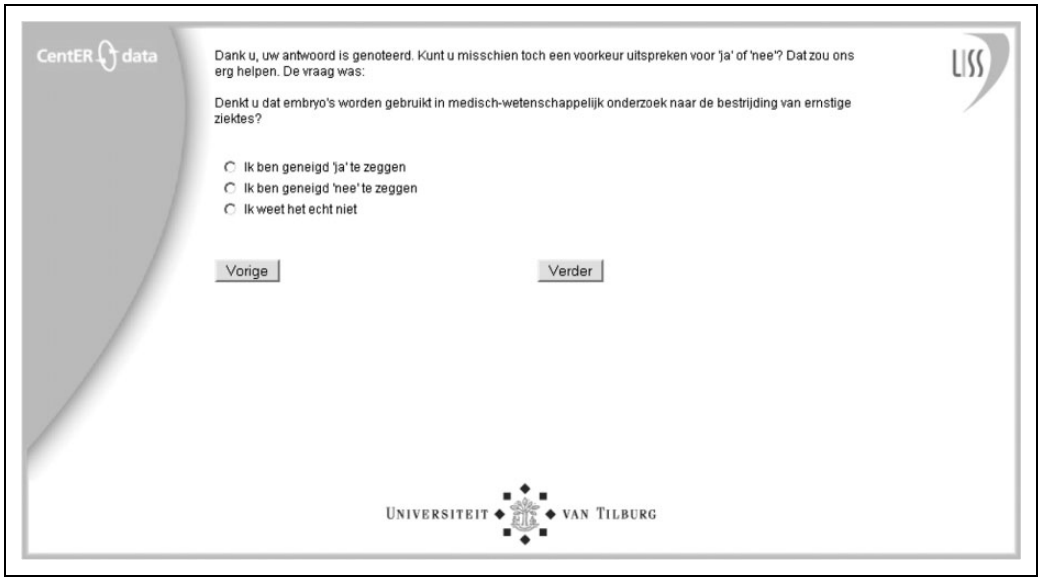
Example follow-up: mandatory response, standard LISS error message: "You have not given an answer"



Example question: do not know ("Weet niet") separated by empty line space



Example question: do not know as a special button "I do not know"



Example probe question: as it appeared online in the conditions skipping allowed, and both forms of explicitly offering do-not-know, see also Appendix A2

Authors' Note

The authors thank Stephanie Stam, Stephanie Mertens, and Annette Scherpenzeel (CentERdata, Tilburg University) and Robert Zandvliet (TNS-NIPO) for their valuable help during the design and data collection. The LISS panel data were collected by CentERdata (Tilburg University, The Netherlands) through its MESS project funded by the Netherlands Organization for Scientific Research (www.lissdata.nl)

Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

References

Al Baghal, T., & Lynn, P. (2014). *Using motivational statements in web instrument design to reduce item missing rates in a mixed-mode context*. (Understanding Society Working Paper Series, 2014-02). Colchester, United Kingdom: Institute for Social and Economic Research, University of Essex.

Behr, D., Kaczmirek, L., Bandilla, W., & Braun, M. (2012). Asking probing questions in web surveys: Which factors have an impact on the quality of responses? *Social Science Computer Review*, *30*, 487–498.

Bethlehem, J., & Biffignandi, S. (2012). *Handbook of web surveys*. New York, NY: John Wiley.

Billiet, J. B., & McClendon, M. J. (2000). Modeling acquiescence in measurement models for two balanced sets of items. *Structural Equation Modeling*, *7*, 608–628.

Converse, J. M. (1976). Predicting don't know in the polls. *Public Opinion Quarterly*, *40*, 515–530.

Converse, J. M., & Presser, S. (1986). *Survey questions: Handcrafting the standard questionnaire*. Beverly Hills, CA: Sage.

Couper, M. P. (2008). *Designing effective web surveys*. Cambridge, MA: Cambridge University Press.

DeRouvray, C., & Couper, M. P. (2002). Designing a strategy for reducing "no opinion" responses in web-based surveys. *Social Science Computer Review*, *20*, 3–9.

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

Dillman, D. A. (2012). Introduction to special issue of survey practice on item nonresponse. *Survey Practice 5*, 2. Retrieved from http://www.surveypractice.org/index.php/SurveyPractice/article/view/44/pdf

Fricker, S., Galesic, M., Tourangeau, R., & Yan, T. (2005). An experimental comparison of web and telephone surveys. *Public Opinion Quarterly*, *69*, 370–392.

Groves, R. M. (1989). *Survey errors and survey costs*. New York, NY: John Wiley.

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

Groves, R. M., & Mc Gonagle, K. A. (2001) A theory-guided interviewer training protocol regarding survey participation. *Journal of Official Statistics*, *17*, 249–265.

Heerwegh, D., & Loosveldt, G. (2008). Face-to-face versus web surveying in a high-internet-coverage population. *Public Opinion Quarterly*, *72*, 836–846.

Hox, J. J. (2010). *Multilevel analysis: Techniques and applications*. New York, NY: Routledge.

Jaeckle, A., Lynn, P., & Burton, J. (2013). *Going online with a face-to-face household panel: Initial results from an experiment on the Understanding Society Innovation Panel*. (Understanding Society Working Paper Series, 2013-03). Colchester, United Kingdom: Institute for Social and Economic Research, University of Essex.

Krosnick, J. A., Holbrook, A. L., Berent, M. K., Carson, R. T., Hanemann, W. M., Kopp, R. J., ... Conaway, M. (2002). The impact of 'don't know' response options on data quality: Nonattitude reduction or an invitation to satisfice? *Public Opinion Quarterly*, *66*, 371–403.

de Leeuw, E. D. (1992). *Data quality in mail, telephone, and face-to-face surveys*. Amsterdam, The Netherlands: TT-publikaties. Retrieved from http://edithl.home.xs4all.nl/pubs/disseddl.pdf

de Leeuw, E. D. (2005). To mix or not to mix data collection modes in surveys. *Journal of Official Statistics*, *21*, 233–255. Retrieved from http://www.jos.nu/Articles/abstract.asp?article=212233

de Leeuw, E. D., & Hox, J. J. (2011). Internet surveys as part of a mixed-mode design. In M. Das, P. Ester, & L. Kaczmirek (Eds.), *Social and behavioral research and the Internet* (pp. 45–76.) EAM methodology series. New York, NY: Routledge.

de Leeuw, E. D., & Hox, J. J. (2014). Survey mode and mode effects. In Engel U., Jann, B., Lynn, P. Scherpenzeel, A. & Sturgis P. (Eds.) Improving survey methods (pp. 22–34). EAM methodology series. New York, NY: Routledge.

de Leeuw, E., Hox, J., & Scherpenzeel, A. (2010). *Emulating interviewers in an online survey: Experimental manipulation over the phone and on the web*. Paper presented at AAPOR 2010, Chicago. Appeared in JSM-proceedings, Survey Methods Section of the American Statistical Association. Retrieved from http://www.amstat.org/sections/SRMS/Proceedings/

de Leeuw, E. D., Hox, J., & Huisman, M. (2003). Prevention and treatment of item nonresponse. *Journal of Official Statistics*, *19*, 153–176. Retrieved from http://www.jos.nu/Articles/abstract.asp?article=192153

Oudejans, M., & Christian, L. M. (2011). Using interactive features to motivate and probe responses to open-ended questions. In M. Das, P. Ester, & L. Kaczmirek (Eds.), *Social and behavioral research and the Internet* (pp. 215–244). New York, NY: Routledge.

Revilla, M. A., & Saris, W. E. (2013). A comparison of the quality of questions in a face-to-face and a web survey. *International Journal of Public Opinion Research*, *25*, 242–253.

Scherpenzeel, A. C., & Das, M. (2011). True longitudinal and probability-based Internet panels: Evidence from the Netherlands. In M. Das, P. Ester, & L. Kaczmirek (Eds), *Social and behavioral research and the Internet* (pp. 77–104) New York, NY: Routledge.

Smyth, J. D., Dillman, D. A., Christian, L. M., & McBride, M. (2009). Open-ended questions in web surveys. Can increasing the size of answer boxes and providing extra verbal instructions improve response quality? *Public Opinion Quarterly*, *73*, 325–327.

Steegers, Ch., Dijstelbloem, H., & Brom, F. W. A. (2008). *Meer dan status alleen. Burgerperspectieven op embryo-onderzoek*. [In Dutch: Assessment and points of view of citizens on technology and embryo research]. The Hague, Netherlands: Rathenau-Instituut, TA rapport 0801. Retrieved from http://www.rathenau.nl/uploads/tx_tferathenau/Meer_dan_status_alleen_april_2008.pdf

Sudman, S., & Bradburn, N.M. (1982). *Asking questions*. San Francisco, CA: Jossey-Bass.

Tourangeau, R., Couper, M. P., & Conrad, F. (2004). Spacing, position, and order. Interpretative heuristics for visual features of survey questions. *Public Opinion Quarterly*, *68*, 368–393.

Tourangeau, R., Conrad, F. G., & Couper, M. P. (2013). *The science of web surveys*. Oxford, England: Oxford University Press.

Wine, J. S., Cominole, M. B., Heuer, R. E., & Riccobono, J. A. (2006). *Challenges of designing and implementing multimode instruments*. Presented at the Second International Conference on Telephone Survey Methodology, Miami, FL, January 2006. Retrieved from https://www.rti.org/pubs/TSM2006_Wine_paper.pdf

Author Biographies

Edith D. de Leeuw is MOA-professor in survey methodology at the department of methodology & statistics, Utrecht University. She is an associate editor of JOS and editor of MDA, and was co-editor of 3 internationally renowned books on survey methodology: The International Handbook of Survey Methodology, Advances in Telephone Methodology, and Survey Measurement and Process Quality. Her research interests focus on mixed-mode and online surveys, children as respondents, total survey error, and comparative and cross-national research. E-mail: e.d.deleeuw@uu.nl

Joop J. Hox (http://joophox.net/) is a professor of social science methodology and statistics at Utrecht University He is the author of Multilevel Analysis, Techniques and Applications and editor of the International Handbook of Survey Methodology. He was the former editor of Methodology and is now president of the European Association of Methodology (EAM). His research interests are data quality in surveys, total survey error, and analysis models for complex data. E-mail: j.hox@uu.nl

Anja Boevé did a combined master in educational sciences and methodology & statistics at Utrecht University. She is a PhD student at the University of Groningen, Heymans Institute department of psychometrics and statistics. Her main research interests are (survey) research methodology, assessment in the field of educational science, and instructional design research. She may be reached at a.j.boeve@rug.nl