Asking probing questions in web surveys: which factors have an impact on the quality of responses?

Behr, Dorothée; Kaczmirek, Lars; Bandilla, Wolfgang; Braun, Michael

Empfohlene Zitierung / Suggested Citation:
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*(4), 487-498. https://doi.org/10.1177/0894439311435305

Social Science Computer Review
30(4) 487-498

DOI: 10.1177/0894439311435305
http://ssc.sagepub.com

Asking Probing Questions in Web Surveys: Which Factors have an Impact on the Quality of Responses?

Dorothée Behr[1], Lars Kaczmirek[1], Wolfgang Bandilla[1], and Michael Braun[1]

Abstract

Cognitive interviewing is a well-established method for evaluating and improving a questionnaire prior to fielding. However, its present implementation brings with it some challenges, notably in terms of small sample sizes or the possibility of interviewer effects. In this study, the authors test web surveys through nonprobability online panels as a *supplemental* means to implement cognitive interviewing techniques. The overall goal is to tackle the above-mentioned challenges. The focus in this article is on methodological features that pave the way for an eventual successful implementation of category-selection probing in web surveys. The study reports on the results of 1,023 respondents from Germany. In order to identify implementation features that lead to a high number of meaningful answers, the authors explore the effects of (1) different panels, (2) different probing variants, and (3) different numbers of preceding probes on answer quality. The overall results suggest that category-selection probing can indeed be implemented in web surveys. Using data from two panels—a community panel where members can actively get involved, for example, by creating their own polls, and a "conventional" panel where answering surveys is the members' only activity—the authors find that high community involvement does not increase the likelihood to answer probes or produce longer statements. Testing three probing variants that differ in wording and provided context, the authors find that presenting the context of the probe (i.e., the probed item and the respondent's answer) produces a higher number of meaningful answers. Finally, the likelihood to answer a probe decreases with the number of preceding probes. However, the word count of those who eventually answer the probes slightly increases with an increasing number of probes.

Keywords

web survey design, probing, open-ended questions, cognitive interviewing, nonprobability online panels

[1]GESIS—Leibniz Institute for the Social Sciences, Mannheim, Germany

Corresponding Author:
Dorothée Behr, GESIS—Leibniz Institute for the Social Sciences, PO Box 12 21 55, 68072 Mannheim, Germany
Email: dorothee.behr@gesis.org

Introduction

Cognitive interviewing is a well-established method for assessing a questionnaire prior to its fielding. However, challenges of cognitive interviewing are equally recognized, such as limited sample sizes or potential interviewer effects due to different probing behavior (e.g., Beatty & Willis, 2007; Blair, Conrad, Ackermann, & Claxton, 2006; Conrad & Blair, 2009). Against this background, we propose to conduct a *supplemental* evaluation activity within web surveys that produces data similar to that produced through cognitive interviews. The overall goal of this new evaluation technique is to circumvent the challenges mentioned above.

The integration of cognitive interviewing techniques into a regular survey is not new. Already in 1966, Schuman pioneered "random probes" in a regular face-to-face survey, and Smith followed his example in 1989. In this article, we aim at preparing the methodological ground for integrating cognitive interviewing techniques, particularly probing, into web surveys. We do so by testing which implementation features contribute to a high number of meaningful probe answers. Further work will need to investigate to what extent these responses can indeed answer substantive research questions.

Web surveys are a very cost- and time-efficient means to conduct studies—both pretesting and postsurvey evaluation studies—with large sample sizes. They allow for meaningful quantification of results as well as analysis of smaller subgroups of respondents with potentially diverging response behavior. They also guarantee standardized probing since each respondent receives the same stimulus. They further permit respondents to take their time to reflect on their answers without feeling pressed, elaborate on their answers or modify them, all of which in complete anonymity. This may in fact speak in favor of probing in web surveys, at least if respondents can be motivated to answer probes in the first place.

The rise of web surveys as a data collection method has led to an array of studies on web design features, such as layout effects, interactivity, or screen aesthetics, and their impact on data quality (Christian, Dillman, & Smyth, 2007; Conrad, Couper, Tourangeau, & Peytchev, 2006; Ganassali, 2008; Mahon-Haft & Dillman, 2010). However, the topic of open-ended questions in web surveys has received attention only recently. The focus to date has been on different answer space sizes (notably for nonnarrative answers), the use of motivational instructions and of follow-ups to open-ended questions or the impact of topic interest and demographic characteristics (Denscombe, 2008; Holland & Christian, 2009; Smyth, Dillman, Christian, & McBride, 2009; Oudejans & Christian, 2010). We advance this research by integrating cognitive interviewing techniques, particularly category-selection probing, into the web environment. Category-selection probing is a standard cognitive interviewing technique. Respondents are asked to name reasons for having selected a particular scale value for a closed question (Prüfer & Rexroth, 2005). Category-selection probing can be used to analyze comprehension problems and the differentiations respondents use in the interpretation of items. Results of category-selection probing can be used to improve questions and to obtain background information on how responses have to be interpreted by analysts.

In this article, we explore the effects of different online panels, different probing variants, and different number of preceding probes on the quality of probe answers. The goal is to identify implementation features that lead to a high number of *productive*, that is, meaningful answers to category-selection probes.

Differences in the composition and practices of online access panels can affect survey results (Baker et al., 2010). A major difference may reside in a community approach where panelists can create their own polls and post their opinions in addition to answering the "usual" surveys. With regard to our research question, we expect community panelists who are particularly active in their panel to be more willing to answer open-ended questions and also to write more words than noncommunity panelists who belong to a "conventional" panel.

To ensure and enhance answer quality among respondents, the probe design should be optimized. Although open-ended questions seem to fare comparably or better in web surveys than in paper-and-pencil surveys and on the whole seem promising (Denscombe, 2008; Holland & Christian, 2009; Smyth et al., 2009), they remain a source of item nonresponse in web surveys. After all, they usually require more effort by respondents than closed-ended questions (Galesic, 2006; Holland & Christian, 2009). Category-selection probing may require even more effort than usual open-ended questions depending on how thoroughly respondents processed and answered the questions that are probed. A suitable probe design should therefore keep respondents' effort at a minimum and at the same time encourage rich and descriptive responses. As to design, this may mean providing the context of the probe (i.e., the closed item and the answer) so that no further effort is needed (such as trying to remember what exactly the question and the answer were or even going back to the previous screen to retrieve the relevant information). In addition, the wording of the probe should prevent respondents from satisficing (Krosnick, 1999). Satisficing, as some argue, may be particularly easy in web surveys where no interviewer is present who could motivate the respondent (Baker et al., 2010). A usual way of asking category-selection probes is to inquire why the respondent has chosen a particular scale value. However, in web surveys—without an interviewer present—the explicit mentioning of the answer category in the probe might invite respondents to find an easy way out of properly answering the probe (e.g., they might only rephrase the given answer category). However, such an answer would not allow us to identify different frames of interpretation. Theoretically, the satisficing risk may be mitigated by deviating attention from the specific answer category and by simply asking respondents to give reasons for their opinion. We test different probe designs and expect that a probe design that reduces respondents' effort by providing the relevant context and that at the same time is less answer value-specific is most successful in eliciting rich and descriptive answers.

Given the effort required to answer open-ended questions, the number of probes across a survey should carefully be chosen. Oudejans and Christian (2010), for instance, show that word count decreases as respondents progress through the survey. We expect that item nonresponse increases and word count decreases, respectively, toward the end of the survey due to increased respondent burden with increasing number of preceding probes.

Method

The data in this article come from two web surveys conducted in Germany in June/July 2010 using two different online panels, henceforth called *community panel* and *noncommunity panel*. The target population was defined in each case as German citizens, aged 18–70. Quotas were set according to region, sex, age, and education.[1] This did certainly not make the survey representative but we tried at least to obtain a broader picture of the population. We targeted 480 completed interviews with the noncommunity panel and 528 with the community panel.[2] The questionnaire was the same for the two panels.

The community panel—where panelists can create their own polls or write opinions—invited their most active community members (i.e., members with the highest polling activity in the last 3 months). We expected that active community members would be more prone to answer open-ended questions and would also produce longer answers than noncommunity panelists.

Furthermore, we designed three variants for category-selection probing (Figures 1–3).

In Variant A, the closed item and the respondent's answer were repeated at the top of the probe screen. The probe itself read: "Please explain why you have chosen [answer category]." Variant B equally had the closed item and the respondent's answer repeated. The wording of the probe, however, was less scale value-specific: "Could you please give reasons for your opinion." In Variant C, neither the closed item nor the answer of the respondent was displayed on the probe screen. Only the

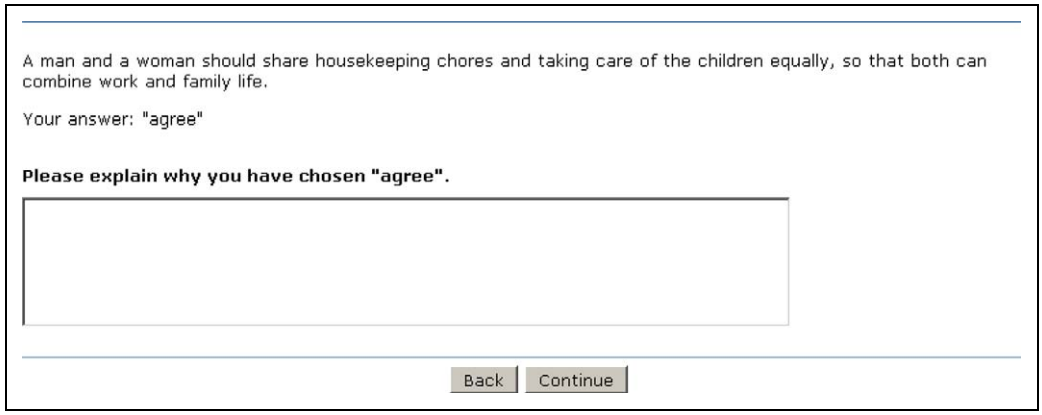

Figure 1. Probe variant A—Closed item and respondent's answer and answer scale-specific probe (item originally in German).

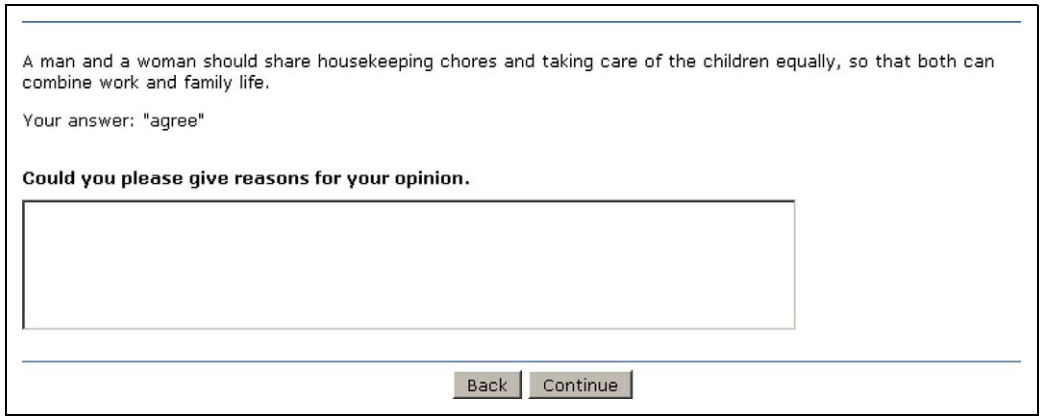

Figure 2. Probe variant B—Closed item and respondent's answer and less answer scale-specific probe (item originally in German).

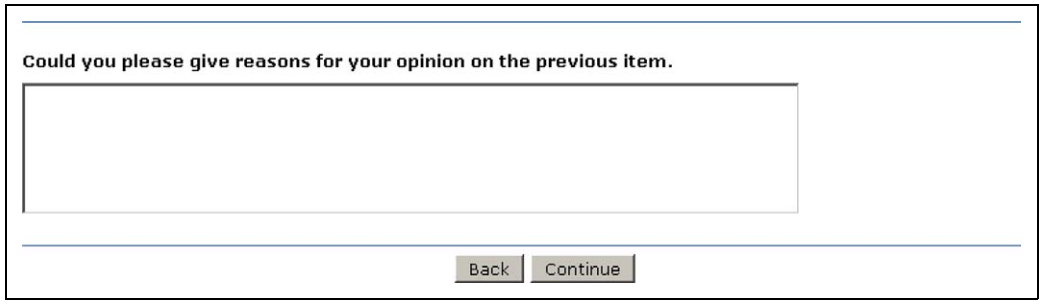

Figure 3. Probe variant C—Probe and no further context (item originally in German).

probe itself was shown: "Could you please give reasons for your opinion on the previous item." We assumed that Probe variant C would be least successful in producing *productive*, that is, meaningful answers due to the missing context on the screen and, therefore, increased respondent burden. In turn, Probe variants A and B would lead to more productive answers because of the provided context. We expected further that the stronger emphasis on the chosen answer category in Variant A might provoke more nonsubstantive answers in terms of just rephrasing the selected answer categories (e.g., because I fully agree) in Variant A than in Variant B. Three agree–disagree items were probed with category-selection probing. These were "A man and a woman should share housekeeping chores and taking care of the children equally, so that both can combine work and family life" (equal division), "A working mother can establish just as warm and secure a relationship with her children as a mother who does not work" (mother–child relationship), and "Having children interferes too much with the freedom of parents" (children constrain freedom). The answers ranged on a 5-point scale from "strongly agree" to "strongly disagree." At the beginning of the survey, the respondents were randomly assigned to Probe variants A, B, or C. This variant then remained constant for them across the survey.

The probed items were part of different topical blocks which were rotated. This allowed us to test whether an increasing number of preceding probes had an effect on the quality of the probe answers—independent of item content. The questionnaire in total comprised 33 closed-ended questions and 6 open-ended questions. This article focuses on the three category-selection probes among the six open-ended questions.

Dependent Variables

To test our assumptions, we used two dependent variables, namely productivity of answers to probes and word count. Productive probe answers were considered to be all answers except the following nonproductive answers: (1) implicit refusals (respondents giving no answer whatsoever), (2) meaningless entries (?, —, or fgh, etc.), (3) don't knows, (4) specified don't knows (respondents providing a reason for their DK such as lack of experience), (5) explicit refusals (respondents answering "n.a.", "no", etc.), (6) other nonsubstantive answers (rhetoric questions such as "why not?", matter-of-fact-statements such as "because it is like that", rephrasing of answer category, etc.), and (7) nonintelligible answers. Examples for (7) are answers such as "equal shares" or "feeling" for the probe after "A working mother can establish just as warm and secure a relationship with her children as a mother who does not work." We are aware that we employed a very broad definition of productivity. Depending on the substantive research question, answers beyond these nonproductive answers may also be unusable. We used a dichotomous measure of productivity (0 = *nonproductive answers*, 1 = *productive answers*) in our models. Analyses with word count, our second dependent variable, were restricted to respondents with at least one productive answer across the three probes.

Independent Variables

Among the independent variables included in our analyses were panel (community panel as the baseline) and two dummies for the probing variants (Variant A constituted the baseline). Control variables were region (western Germany as the baseline), sex (women as the baseline), geographical origin (German origin as the baseline), education (university entrance requirement as the baseline), and year of birth. Furthermore, the item "A man's job is to earn money; a woman's job is to look after the home and family" which measures traditionality with regard to gender roles was used as an attitudinal control variable. Its answers on a 5-point scale ranged from "strongly agree" to "strongly disagree." The number of preceding probes was operationalized by a quantitative variable, taking on the values 0–5.

Table 1. Item- and Panel-Specific Outcomes in Productive Answers and Word Count

	Productive Probe Answers (%)	Word Count of Probe Answers
Item "Equal division"		
Noncommunity panel	83.71	1–152 (mean: 21.1)
Community panel	80.86	1–162 (mean: 19.0)
Total	82.21	1–162 (mean: 20.0)
Item "Mother–child relationship"		
Noncommunity panel	78.76	1–103 (mean: 23.3)
Community panel	68.40	1–136 (mean: 20.4)
Total	73.31	1–136 (mean: 21.9)
Item "Children constrain freedom"		
Noncommunity panel	80.00	1–116 (mean: 19.5)
Community panel	78.44	1–189 (mean: 17.9)
Total	79.18	1–189 (mean: 18.7)

Note. $N = 1{,}023$, but word count excludes nonproductive answers.

Analytical Procedure

Since the answers to the probes were not independent of each other but nested in respondents, we employed multilevel modeling. Multilevel modeling is appropriate when variables pertain to different levels and when a dependency exists between the elements of the lower level. In our models, the lower level is constituted by the three probes, the higher level by the respondents. We estimated separate multilevel models for productivity and for word count, using an identical set of independent variables.

We started with an empty model containing productivity as our dependent variable and also a differentiation between the three probed items (Model 0). This model allowed us to decompose the variance between the probe level and respondent level and to assess the appropriateness of multilevel modeling. Model 1 included the respondent-level characteristics panel and probe variants. In Model 2, we added the control variables region, sex, geographical origin, education, and year of birth. Model 3 included the attitudinal item as an additional control variable. Model 4 included the number of preceding probes. The hierarchy in models was determined by first introducing the higher-level variables, including the control variables pertaining to this level, and then adding the preceding probe number as our only lower-level variable. The models for the word count followed the same logic. Likelihood-ratio tests were performed to establish difference in fit between the models.

Results

485 noncommunity panel members completed the survey; the drop-out rate was 7.8% (41/526). The mean completion time was 14:38 min (median 11:34 min.). From the community panel, 538 members completed the survey; here the drop-out rate was 6.4% (37/575). The mean completion time for the community panel was 12:54 min (median 10:00 min.). Across the panels and across the 3 probed items, between 73.31% and 82.21% of probe answers were coded as productive. Interrater reliability based on two independent codings and 10% of answers for each probed item ranged from 90% to 98%.[3] The average word count of productive answers ranged from 18.7 to 21.9 words. Table 1 presents results separately for probed items and panels.

With multilevel modeling of productivity (Table 2), we find the following: Model 0 (not presented in detail) shows that 71% of the total variance in productivity derives from respondent differences and the remainder of 29% from differences on the probe level. This distribution justifies the use of multilevel modeling. Contrary to our expectations, Model 1 shows that noncommunity

Table 2. Mixed-Effects Logistic Regression Models Predicting Productive Answer Behavior (Odds Ratios, in Parentheses: z Values)

	Model 0	Model 1	Model 2	Model 3	Model 4
Probed items (base: equal division)					
Mother–child relationship	.342*** (−6.70)	.342*** (−6.70)	.340*** (−6.72)	.340*** (−6.72)	.382 (−5.80)***
Children constrain freedom	.659** (−2.60)	.659** (−2.60)	.657** (−2.61)	.657** (−2.62)	.700 (−2.18)*
Panel: noncommunity panel (base: community panel)	–	1.765* (2.35)	1.782* (2.46)	1.547 (1.85)	1.532 (1.80)
Probe variants (base: A)					
Probe variant B		.869 (−0.47)	.805 (−0.75)	.803 (−0.76)	.800 (−0.77)
Probe variant C		.529* (−2.16)	.519* (−2.29)	.533* (−2.21)	.524 (−2.25)*
Sociodemographic variables					
Region: eastern Germany (base: western Germany)	–	–	1.213 (0.83)	1.113 (0.46)	1.116 (0.47)
Sex: men (base: women)	–	–	.186*** (−6.84)	.198*** (−6.64)	.192 (−6.68)***
Geographical origin: non-German origin (base: German origin)	–	–	.610 (−0.76)	.634 (−0.70)	.642 (−0.67)
Year of birth	–	–	.959*** (−4.93)	.962*** (−4.55)	.962 (−4.46)***
Education: less than university entrance requirement (base: university entrance requirement)			.717 (−1.43)	.812 (−0.89)	.814 (−0.87)
Attitudinal item	–	–	–	1.490*** (3.77)	1.502 (3.81)***
No. of preceding probes (0–5)	–	–	–	–	.883 (−2.80)**

Note. $N = 1{,}007$.
*$p < .05$. **$p < .01$. ***$p < .001$.

panelists are more likely to produce productive answers. In addition, Probe variant C, where no further context is provided on the screen, is less likely than Probe variant A, our baseline, to produce productive answers. No differences can be found between Probe variants A and B, though. In Model 2, we control for the sociodemographic background of the respondents. The panel and probe variant effects remain significant. In addition, we find that women and older respondents are more likely to produce a productive answer than men and younger respondents. Education, however, does not impact on the likelihood to produce a productive answer. Upon introducing the attitudinal benchmark item in Model 3, the difference between the panels disappears while the effects of probe variants, sex, and age remain significant. We find that nontraditional respondents are more likely to write productive answers than traditional respondents. Obviously, nontraditional attitudes are related to a higher interest in the topic of the questions. The introduction of the preceding probe number in Model 4 does not alter the conclusions so far. It shows, however, that with increasing number of probes productive answers are less likely. The improvements across all models are statistically significant following likelihood-ratio tests ($p < .05$).

With the dependent variable word count (Table 3), we follow the same approach as above in building our models. Our Model 0 decomposes the variance in word count between probe level and respondent level: the respondents account for 58% of variance, the probes for the remaining 42%. In Model 1, we include the probe variants as well as the panel. We find that the noncommunity panelists write more words than the community panelists. The two-word difference between the panels, though, is rather minimal. The probe variant has no impact on the word count as nonproductive answers were not included in the analyses. In Model 2, sociodemographic background variables are added. Women write significantly more words than men; at the same time, increasing age and particularly lower education is related with writing fewer words. Model 3 shows that nontraditional respondents write more words than traditional respondents. When controlling for traditionality, the panel effect found in Model 1, namely that noncommunity panelists write more, weakens although it remains significant. Finally, Model 4 includes the number of preceding probes. With increasing number of probes, respondents seem to write more, although effects are certainly minimal. Among productive respondents at least, we are thus observing a warming-up effect. Likelihood-ratio tests show that the improvements across all models are statistically significant ($p < .05$).

Discussion

Overall, it seems that category-selection probing can successfully be implemented in web surveys for which respondents are drawn from online panels. On the whole, respondents provide productive answers to category-selection probes. In this article, we investigated which implementation features need to be considered in order to obtain a high number of productive probe answers.

Contrary to our assumptions, the panel character *per se*—whether community or noncommunity panel—does not or hardly influence the answer quality. We rather conclude that different panel distributions on a key attitude that is related to the topic of the probes and can be taken as an indicator for interest have an impact on both the likelihood to respond and on word count. On this key attitude, namely the attitudinal item introduced into the third model, the panels differed: the noncommunity panel was more egalitarian in gender role attitudes than the community panel. Different distributions on attitude items can, therefore, make a difference between panels when it comes to answering probe items. The panel composition thus becomes an important factor to take into account when choosing a panel provider for a probing study. In line with one's research question, a researcher may choose a panel with panel members that predominantly harbor certain attitudes or certain interests (if this is known in advance). Other research questions may call for a panel whose coverage of the general (Internet) population is as broad as possible. A certain level of nonresponse on probe answers among those less interested in a topic will then have to be accepted, however. Regardless of the panel

Table 3. Mixed-Effects Maximum Likelihood (ML) Regression Predicting Word Count (Coefficients, in Parentheses: z Values)

	Model 0	Model 1	Model 2	Model 3	Model 4
Probed items (base: equal division)					
Mother–child relationship	1.152 (1.90)	1.140 (1.88)	1.110 (1.84)	1.117 (1.85)	0.730 (1.16)
Children constrain freedom	−1.662** (−2.80)	−1.657** (−2.79)	−1.634** (−2.75)	−1.631** (−2.75)	−1.817 (−3.03)**
Panel: noncommunity panel (base: community panel)	—	2.384* (2.32)	2.448* (2.44)	2.006* (1.98)	2.021 (2.00)*
Probe variants (base: A)					
Probe variant B	—	2.201 (1.76)	2.057 (1.68)	2.028 (1.67)	2.044 (1.68)
Probe variant C	—	.915 (0.72)	.637 (0.51)	.672 (0.54)	0.707 (0.57)
Sociodemographic variables					
Region: eastern Germany (base: western Germany)	—	—	−.006 (−0.01)	−.259 (−0.26)	.283 (0.28)
Sex: men (base: women)	—	—	−5.803*** (−5.78)	−5.550*** (−5.52)	−5.513 (−5.49)***
Geographical origin: non-German origin (base: German origin)	—	—	1.904 (0.64)	1.899 (0.64)	1.841 (0.62)
Year of birth			0.084* (2.37)	0.093** (2.60)	0.092 (2.58)*
Education: less than university entrance requirement (base: university entrance requirement)			− 3.702*** (−3.68)	−3.341** (−3.30)	−3.339 (−3.30)**
Attitudinal item	—	—	—	1.278** (2.67)	1.272 (2.66)**
No. of preceding probes (0–5)	—	—	—	—	.383 (2.19)*

Note. $N = 910$.
*$p < .05$. **$p < .01$. ***$p < .001$.

composition, effects of sex, age, or education seem unavoidable. Here, the study backs findings from Oudejans and Christian (2010) as well as Denscombe (2008).

A probe design which had the closed item, the respondent's answer to it, and the probe on a screen was most successful in eliciting productive answers compared to just the probe without any context (Variant C). Further differentiation in probe wording (Variant A or B) did not produce significantly different results. However, for those who eventually answered the probe, the probe variants did not have an effect on the word count. Efforts should, therefore, be made toward providing respondents with the needed context on the screen in order to reduce nonproductive answers to category-selection probes.

In terms of number of probes, we found that the probes had not yet imposed too heavy a burden on the respondents. The odds of responding in a productive manner decreased with an increasing number of preceding probes. At the same time—among productive respondents—the writing level did not abate, rather the contrary. On the whole, keeping up the motivation and integrating probes sensibly and sparingly is thus an important issue for future studies.

This article provides us with positive results on the willingness of online panelists to answer category-selection probes on the web. We are now more aware of overall design features or respondent characteristics that lead to providing answers to these questions. The steps currently taken are to estimate the usefulness of answers received with regard to answering substantive research questions. Behr, Braun, Kaczmirek, and Bandilla (in press), for instance, show that online probe answers help to uncover validity problems with a gender ideology item that is meant to measure a nontraditional stance but falls short of this goal. Further studies currently undertaken involve online probing in the cross-national context as well as the inclusion of other probe types besides category-selection probing. By asking a specific probe, Braun, Behr, and Kaczmirek (2011), for instance, explore what type of immigrants respondent in different countries have in mind when answering attitude items on immigrants. While the immigrant groups mentioned differ across countries, on an abstract level they seem to be comparable.

Despite our satisfaction with the overall results of online probing, we caution against substituting this online tool for face-to-face cognitive interviewing. We rather envisage it as a *supplemental* tool for situations where quantification of results and the coverage of hard-to-reach population groups, for example, those identified by contradictory answer behavior, are needed.

We wish to stress that the web implementation cannot offer targeted follow-up probes to incomprehensible or insufficient probe answers nor can it ask respondents to elaborate on an issue if the researcher feels this would be necessary. This interactivity is impossible, unless it is known in advance what exactly constitutes an incomprehensible or insufficient answer. However, the possibilities to add a follow-up probe to nonproductive probe answers, such as "???" or "don't know," are available, and also pioneered in research (e.g., Oudejans & Christian, 2010). This may be a research line worth taking up in order to enhance the chances of getting at least a minimal response to a probe question. Further research could also investigate general differences between face-to-face and online probing as well as the use of different probe types in web surveys.

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 disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was funded by the German Research Foundation (DFG) as part of the PPSM Priority Programme on Survey Methodology (SPP 1292) (project # BR 908/3-1).

Notes

1. Region included eastern and western Germany. Age groups were defined as 18–30, 31–50, and 51–70. Education differentiated between higher secondary education (university entrance requirement) and lower or no education (less than university entrance requirement). An equal number of cases was targeted for each combination.
2. Learning from the first panel survey, a 10% increase was implemented for the survey with the second panel to compensate for lurkers who quickly clicked through the survey without giving any useful answers to probe questions.
3. Discrepancies almost exclusively pertained to the category "nonintelligible." If coders had had a particular substantive research question in mind, these discrepancies would possibly have been reduced.

References

Baker, R., Blumberg, S. J., Brick, J. M., Couper, M. P., Courtright, M., Dennis, J. M., ... Zahs, D. (2010). AAPOR report on online panels. *Public Opinion Quarterly*, *74*, 711–781.

Beatty, P. C., & Willis, G. (2007). Research synthesis: The practice of cognitive interviewing. *Public Opinion Quarterly, 71*, 287–311.

Behr, D., Braun, M., Kaczmirek, L., & Bandilla, W. (in press). Testing the validity of gender ideology items by implementing probing questions in web surveys. *Field Methods*.

Blair, J., Conrad, F., Ackermann, A. C., & Claxton G. (2006). *The effect of sample size on cognitive interview findings*. Paper presented at the AAPOR Conference, Montreal, Canada, May 18–21, 2006. Retrieved August 9, 2011, from http://www.abtassociates.com/presentations/aapor06_sample_size_cognitive_interviews.pdf

Braun, M., Behr, D., & Kaczmirek, L. (2011). *Assessing cross-national equivalence of measures of xenophobia: Evidence from probing in web surveys*. (under review).

Christian, L. M., Dillman, D. A., & Smyth J. D. (2007). Helping respondents get it right the first time: The influence of words, symbols, and graphics in web surveys. *Public Opinion Quarterly, 71*, 113–125.

Conrad, F. G., & Blair, J. (2009). Sources of error in cognitive interviews. *Public Opinion Quarterly, 73*, 32–55.

Conrad, F. G., Couper, M. P., Tourangeau, R., & Peytchev, A. (2006). Use and non-use of clarification features in web surveys. *Journal of Official Statistics, 22*, 245–269.

Denscombe, M. (2008). The length of responses to open-ended questions: A comparison of online and paper questionnaires in terms of a mode effect. *Social Science Computer Review, 26*, 359–368.

Galesic, M. (2006). Dropouts on the Web: Effects of interest and burden experienced during an online survey. *Journal of Official Statistics, 22*, 313–328.

Ganassali, S. (2008). The influence of the design of web survey questionnaires on the quality of responses. *Survey Research Methods, 2*, 21–32.

Holland, J. L., & Christian, L. M. (2009). The influence of topic interest and interactive probing on responses to open-ended questions in web surveys. *Social Science Computer Review, 27*, 196–212.

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

Mahon-Haft, T. A., & Dillman, D. A. (2010). Does visual appeal matter? Effects of web survey aesthetics on survey quality. *Survey Research Methods, 4*, 43–59.

Oudejans, M., & Christian, L. M. (2010). 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: Advances in applied methods and research strategies* (pp. 304–332). London; NY: Routledge.

Prüfer, P., & Rexroth, M. 2005. Kognitive interviews [cognitive interviews]. ZUMA How-to-Reihe 15. Retrieved August 2, 2011, from http://www.gesis.org/fileadmin/upload/forschung/publikationen/gesis_reihen/howto/How_to15PP_MR.pdf?download=true

Schuman, H. 1966. The random probe: A technique for evaluating the validity of closed questions. *American Sociological Review, 31*, 218–222.

Smith, T. W. 1989. Random probes of GSS questions. *International Journal of Public Opinion Research, 1*, 305–325.

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*, 325–337.

Bios

Dorothée Behr is a senior researcher in the department of survey design and methodology at GESIS—Leibniz Institute for the Social Sciences, Mannheim (Germany). She holds a doctoral degree from the University of Mainz (2009). Her current research interests include web survey design and comparability of cross-cultural questionnaires. E-mail contact: dorothee.behr@gesis.org.

Lars Kaczmirek is a senior researcher at GESIS—Leibniz Institute for the Social Sciences in Mannheim (Germany). He received his PhD in psychology and specialized in survey design and methodology. His publications focus on online survey methodology, data quality, eyetracking research, data protection, accessibility, and usability. E-mail contact: lars.kaczmirek@gesis.org.

Wolfgang Bandilla is a project consultant and senior researcher in the department of survey design and methodology at GESIS—Leibniz Institute for the Social Sciences, Mannheim (Germany). He has specialized in web survey methodology and mixed-mode surveys. E-mail contact: wolfgang.bandilla@gesis.org.

Michael Braun is a senior project consultant at GESIS—Leibniz Institute for the Social Sciences and adjunct professor at the University of Mannheim (Germany). He has specialized in cross-cultural survey methodology and analysis. He has extensively published both on methodological problems of interculturally comparative research and on international comparisons in the fields of migration, work, and the family. E-mail contact: michael.braun@gesis.org.