Open Access Repository
www.ssoar.info

Effects of the Number of Open-Ended Probing Questions on Response Quality in Cognitive Online Pretests

Neuert, Cornelia; Lenzner, Timo

Veröffentlichungsversion / Published Version
Zeitschriftenartikel / journal article

Zur Verfügung gestellt in Kooperation mit / provided in cooperation with:
GESIS - Leibniz-Institut für Sozialwissenschaften

Empfohlene Zitierung / Suggested Citation:
Neuert, C., & Lenzner, T. (2021). Effects of the Number of Open-Ended Probing Questions on Response Quality in Cognitive Online Pretests. *Social Science Computer Review*, *39*(3), 456-468. https://doi.org/10.1177/0894439319866397

Nutzungsbedingungen:
Dieser Text wird unter einer Deposit-Lizenz (Keine Weiterverbreitung - keine Bearbeitung) zur Verfügung gestellt. Gewährt wird ein nicht exklusives, nicht übertragbares, persönliches und beschränktes Recht auf Nutzung dieses Dokuments. Dieses Dokument ist ausschließlich für den persönlichen, nicht-kommerziellen Gebrauch bestimmt. Auf sämtlichen Kopien dieses Dokuments müssen alle Urheberrechtshinweise und sonstigen Hinweise auf gesetzlichen Schutz beibehalten werden. Sie dürfen dieses Dokument nicht in irgendeiner Weise abändern, noch dürfen Sie dieses Dokument für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, aufführen, vertreiben oder anderweitig nutzen.
Mit der Verwendung dieses Dokuments erkennen Sie die Nutzungsbedingungen an.

Terms of use:
This document is made available under Deposit Licence (No Redistribution - no modifications). We grant a non-exclusive, non-transferable, individual and limited right to using this document. This document is solely intended for your personal, non-commercial use. All of the copies of this documents must retain all copyright information and other information regarding legal protection. You are not allowed to alter this document in any way, to copy it for public or commercial purposes, to exhibit the document in public, to perform, distribute or otherwise use the document in public.
By using this particular document, you accept the above-stated conditions of use.

gesis
Leibniz-Institut
für Sozialwissenschaften

Mitglied der
Leibniz-Gemeinschaft

Diese Version ist zitierbar unter / This version is citable under:
https://nbn-resolving.org/urn:nbn:de:0168-ssoar-85353-1

Social Science Computer Review
2021, Vol. 39(3) 456-468
© The Author(s) 2019
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/0894439319866397
journals.sagepub.com/home/ssc
SAGE

Effects of the Number of Open-Ended Probing Questions on Response Quality in Cognitive Online Pretests

Cornelia E. Neuert[1] **and Timo Lenzner**[1]

Abstract

Cognitive online pretests have, in recent years, become recognized as a promising tool for evaluating questions prior to their use in actual surveys. While existing research has shown that cognitive online pretests produce similar results to face-to-face cognitive interviews with regard to the problems detected and the item revisions suggested, little is known about the ideal design of a cognitive online pretest. This study examines whether the number of open-ended probing questions asked during a cognitive online pretest has an effect on the quality and depth of respondents' answers as well as on respondents' satisfaction with the survey. We conducted an experiment in which we varied the number of open-ended probing questions that respondents received during a cognitive online pretest. The questionnaire consisted of 26 survey questions, and respondents received either 13 probing questions ($n = 120$, short version) or 21 probing questions ($n = 120$, long version). The findings suggest that asking a greater number of open-ended probes in a cognitive online pretest does not undermine the quality of respondents' answers represented by the response quality indicators: (1) amount of probe nonresponse, (2) number of uninterpretable answers, (3) number of dropouts, (4) number of words, (5) response times, and (6) number and type of themes covered by the probes. Furthermore, the respondents' satisfaction with the survey is not affected by the number of probes being asked.

Keywords

open-ended questions, probes, online pretest, web probing, questionnaire design, web survey

Cognitive online pretests have, in recent years, become recognized as a promising tool for evaluating questions prior to their use in actual surveys (Lenzner & Neuert, 2017). The term cognitive online pretests or web probing[1] refers to the implementation of mainly open-ended but also closed-ended probing questions in online questionnaires. These probes are adopted from cognitive interviewing and, as with face-to-face cognitive interviewing, are intended to gather information about the

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

Corresponding Author:
Cornelia E. Neuert, GESIS—Leibniz Institute for the Social Sciences, P.O. Box 12 21 55, 68072 Mannheim, Germany.
Email: cornelia.neuert@gesis.org

response process. Cognitive probing questions thus provide information on the validity of survey questions and can be used to determine whether a respondent understands the meaning of a question as intended by the researcher and whether all respondents interpret a question or term in the same way. Compared to traditional face-to-face cognitive interviews, one advantage of web probing is that it enables the quick and cost-effective recruitment of participants. It is thus much easier to realize large sample sizes and to recruit participants from different geographic regions in comparison to face-to-face cognitive interviewing. In addition, increased standardization can be achieved when implementing probing questions in a web survey because the administration of the probing questions is determined from the outset. And finally, the fact that the cognitive techniques employed in an online survey are self-administered diminishes potential interviewer effects (Behr, Bandilla, Kaczmirek, & Braun, 2014; Lenzner & Neuert, 2017; Meitinger & Behr, 2016).

These advantages are offset by the fact that probes must be developed and programmed in advance, which makes the method less flexible (i.e., there is no way to follow up on responses that are very short and thus difficult to interpret). Moreover, due to the absence of an interviewer, the motivating effect of the interviewer to answer the open-ended questions (satisfactorily) is missing.

Nevertheless, past research has shown that cognitive online pretests produce similar results to face-to-face cognitive interviews with regard to the problems detected and the item revisions suggested (Lenzner & Neuert, 2017; Meitinger & Behr, 2016). Despite these promising research findings, until now little has been established regarding the ideal design of a cognitive online pretest (e.g., the ideal length of the pretest or the maximum number of open-ended probing questions that respondents are able or willing to answer). In this study, we therefore address one of these research gaps by examining whether the number of open-ended probing questions asked during a cognitive online pretest has an effect on the quality and depth of respondents' answers. This enables us to evaluate how response quality changes with an increasing number of preceding probes in order to make recommendations for the practical implementation of probes in the context of cognitive online pretests.

Background

As early as the mid-1960s, Schuman (1966) was aware of the advantages of implementing open-ended questions in surveys. In "The Random Probe" (1966), he argued that probing a randomly selected subset of responses to closed-ended survey questions could provide insights into the basis of the response or reveal potential need for clarification. However, the literature on open-ended questions has established that answering these questions places a greater burden on respondents' cognitive abilities than selecting a response category in closed-ended questions, since respondents must formulate the answer in their own words and express it verbally or in writing (Dillman, Smyth, & Christian, 2009; Züll, 2016). Given the effort required, Dillman, Smyth, and Christian (2009) generally recommend using open-ended questions sparingly to avoid overburdening respondents and to encourage their willingness to respond. Due to the more cognitively demanding answer process for open-ended questions, responses to open-ended questions are often prone to higher rates of item nonresponse (Borg & Zuell, 2012; Denscombe, 2008; Reja, Manfreda, Hlebec, & Vehovar, 2003; Scholz & Zuell, 2012; Züll, Menold, & Körber, 2014). Although the amount of item nonresponse on open-ended questions is lower in web surveys than in paper-and-pencil surveys (Denscombe, 2008; Kwak & Radler, 2002), current research on web probing shows that the proportion of noninterpretable answers (e.g., meaningless letter combinations such as "asdf") and unanswered questions is significantly higher in web probing than in face-to-face cognitive interviewing (Meitinger & Behr, 2016). According to Behr, Kaczmirek, Bandilla, and Braun (2012), answering cognitive probing questions may place even more burden on respondents than common open-ended questions, depending on how thoroughly respondents process and answer the questions that are followed up on.

With regard to the number of open-ended questions, Oudejans and Christian (2010) show that the length of the answers decreases with an increasing number of preceding open-ended questions. In the context of web probing, Behr et al. (2012) note that the likelihood of respondents producing productive answers to probes decreases with an increasing number of probes. However, for some respondents, a "warming-up effect" was observable, meaning that these respondents gave slightly longer answers the more probes they received. This finding suggests that besides being potentially perceived as burdensome, giving respondents the opportunity to comment on the key topics covered in the survey and to voice their own views can also lead to higher respondent engagement and satisfaction with the survey. Unfortunately, respondents in the study by Behr et al. (2012) only received a maximum of six probes so that the total number of preceding probing questions was limited to five. We are not aware of any research to date examining the effects of a higher number of probes on response quality and respondent satisfaction in web probing studies.

Research Objectives

In this article, we address the following research questions:

Research Question 1: Does the number of open-ended probing questions asked during a cognitive online pretest have an effect on the quality of respondents' answers?

Research Question 2: Does the number of open-ended probing questions have an effect on respondents' satisfaction with the survey?

Research Question 3: Does the number of open-ended probing questions asked during a cognitive online pretest have an effect on the content or depth of respondents' answers?

To answer these research questions, we systematically varied the number of open-ended probes in a cognitive online pretest resulting in two experimental groups ("short" vs. "long," with 13 vs. 21 open-ended probes, respectively). To examine Research Question 1, we used the following response quality indicators to compare the responses of the 13 probes that were asked in both questionnaire versions: (1) amount of probe nonresponse, (2) number of uninterpretable answers, (3) number of dropouts, (4) number of words respondents type in per probe, and (5) response times. Assuming that a higher number of open-ended probes add to response burden and thereby decreases response quality, we expect the amount of item nonresponse, uninterpretable answers, and dropouts being higher in the long condition than in the short and average word count and response times for the probes to be lower or shorter in the long condition than in the short (Hypothesis 1).

To answer Research Question 2, we implemented a question at the end of the questionnaire asking respondents to rate their satisfaction with the survey on a 5-point scale ranging from *very poor* to *very good*. With regard to respondents' satisfaction with the questionnaire, we do not have an overarching, directed hypothesis because two scenarios are conceivable: On the one hand, it is plausible that people will evaluate the many open-ended probes positively because they can express and justify their opinions more freely. On the other hand, it is also possible that respondents will find the high number of probes annoying and burdensome and therefore rate the survey less positively.

To examine Research Question 3, we compare how many topics are covered by the open-ended questions and whether this differs with regard to the number of probes asked in total. Overall, we expect the number of topics covered to be lower in the long condition than in the short condition (Hypothesis 2).

Methods and Data

The questionnaire originated from a pretest project that was carried out by one of the authors in 2015 on "Subjective Feeling of Security in One's Neighborhood" (Lenzner, Disch, Gebhardt, & Menold,

Table 1. Total Number of Probes Per Question Across Conditions.

	Number of Probes Per Condition						
		Question					
Condition	Total	1	2	3	4	5	6
Probes							
Long (N = 120)	21	1	5	5	5	4	1
Short (N = 120)	13	1	3	3	3	2	1

Table 2. Overview of the Distribution of Probing Questions Across Questions and of the Probe Position Within the Questionnaire.

Question	1	2					3				
Probe	P1_1	P2_1	P2_2	P2_3	P2_4	P2_5	P3_1	P3_2	P3_3	P3_4	P3_5
Probe number (short/long)	1/1	2/2	3/3	—/4	—/5	4/6	5/7	6/8	—/9	—/10	7/11

Question	4					5				6
Probe	P4_1	P4_2	P4_3	P4_4	P4_5	P5_1	P5_2	P5_3	P5_4	P6_1
Probe number (short/long)	8/12	—/13	9/14	10/15	—/16	11/17	—/18	12/19	—/20	13/21

Note. The table lists all probes in the long condition and shows whether this question was asked in the short condition or not, and the position of the probe within the questionnaire (e.g., P3_4 is the fourth probe following Question 3 and was the tenth probe in the long condition, while it was not asked in the short condition).

2015). The field time for this experiment was November 2017. The questionnaire consisted of 26 survey questions in German (two single questions and four grid questions with 6 items, respectively). In this study, we varied whether respondents received 13 probing questions ($N = 120$, short version) or 21 probing questions ($N = 120$, long version; see Table 1). The number of survey questions and the sequence in both conditions were kept constant so that the quality of the answers could be compared both within the individual surveys and between the surveys (for 13 open-ended questions). A detailed overview of the distribution of probing questions across questions and of the probe position within the questionnaire is shown in Table 2.

Several probing techniques were used in the online survey: general/elaborate probes (i.e., "Can you explain your answer in more detail?"), comprehension probes (i.e., "What do you understand by 'your immediate neighborhood'?"), and specific probes (i.e., "Which public 'green areas' did you have in mind when answering the question?"). Probing questions were presented on separate screens following the respective survey questions (except for probe 9/14 and 10/15 which were presented below each other on one screen). In addition to the probe, the corresponding question text and, if relevant, the provided answer were also displayed on the probe screen.

The respondents were drawn from a German nonprobability online panel using cross-quotas for age, education, and gender. We aimed for a net sample size of 120 completed questionnaires in each condition. Table 3 shows the key demographic characteristics of the two experimental groups.

Respondents were allowed to participate via any computer device (desktop PC, tablet, mobile phone). The amount of respondents participating via mobile devices were comparable across conditions, with 23% in the short and 29% in the long condition ($\chi^2 = 1.392$, $p = .238$). After the

Table 3. Demographic Characteristics of Respondents.

Condition	Short (%)	Long (%)	χ^2	*p*
Sex				
Female	50 (42)	60 (50)	1.678	.195
Male	70 (58)	60 (50)		
Age				
18–30	37 (31)	34 (28)	.333	.846
31–50	44 (37)	43 (36)		
51–70	39 (32)	43 (36)		
Education				
Less than college	72 (60)	71 (59)	.017	.895
College and higher	48 (40)	49 (41)		
N	120	120		

welcome page of the online survey, all respondents were informed that the web survey was a pretest of the questionnaire and that the questions would be revised based on their comments to the open-ended probing questions. The announced completion time was 15–20 min for the short and 20–25 min for the long condition. The actual average completion time of the survey was 14 min in the short version and 22 min in the long version. Depending on the announced completion time, respondents received an incentive for participation according to the incentive system of the online panel provider, which was slightly higher in the long version (€2.00) than in the short version (€1.50).

Results

Response Quality and Respondent Satisfaction

We examined five measures of response quality for the 13 overlapping probes: (1) amount of probe nonresponse, (2) number of uninterpretable answers, (3) and number of dropouts, (4) average word count for the open-ended probes, and (5) average response times. In order to reduce the number of statistical tests and to efficiently summarize the results, we conducted the analyses for the quality indicators' word count and response times on the means of the open-ended probes and for the remaining three quality indicators on the proportion of respondents showing these response behaviors. Respondents' satisfaction with the questionnaire was measured by a question at the end of the questionnaire asking respondents to rate their satisfaction with the survey on a 5-point scale ranging from *very poor* to *very good*. The results for the five indicators are shown in Table 4.

Amount of probe nonresponse. We differentiate between three types of probe nonresponse: (1) complete nonresponse, which refers to respondents who skipped the probe question and stated "I don't want to give any information here" by checking an answer box that appeared at the top of the survey page after they had clicked the submit button without providing an answer; (2) respondents who explicitly refused to answer (e.g., "—," "no,") or who gave no usable responses ("fgsgdg"); and (3) respondents who expressed uncertainty or insufficient knowledge (e.g., "I don't know," "?").

The overall amount of probe nonresponse did not differ across the two conditions and was approximately 20% in both versions (χ^2 = .40, p = .528). Nonresponse per probe varied between 12% and 24% in the short version and between 11% and 28% in the long version. Figure 1 shows how the probe nonresponse rates develop over the course of the questionnaire. In both versions, the amount of nonresponse increases over time and from probe 8/12 onward, it is consistently higher in the long version than in the short version (except for probe 12/19). However, the analyses of χ^2 tests

Table 4. Means of the Response Quality Indicators Word Count and Response Times, and Proportion of Response Quality Indicators Item Nonresponse, Uninterpretable Answers, and Dropouts.

	Short		Long		Test	
Response Quality Indicators	%	*N*	%	N	χ^2	*p*
Item nonresponse overall	19.4	300	20.3	314	.398	.528
Complete nonresponse	4.7	73	5.6	86		
Refusal/not useful	13.1	202	12.1	187		
Don't know	1.6	25	2.7	41		
Uninterpretable answers	11.4	142	10.6	130	.278	.598
Dropouts	18.4	27	13.0	18	1.517	.218
	M	*SD*	*M*	*SD*	*t*	*p*
Word count	7.5	5.9	8.6	7.5	−1.344	.180
Response times	37.7	31.4	42.6	31.7	−1.200	.231

Note. M = mean; *SD* = standard deviation; short = 13 open-ended probes; long = 21 open-ended probes.

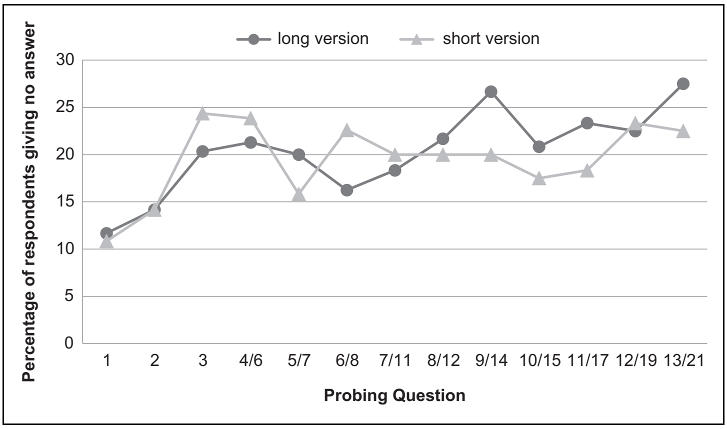

Figure 1. Amount of nonresponse (in %) per probing question.

showed no significant differences in the amount of probe nonresponse between the two versions. Hypothesis 1, stating that the amount of nonresponse in the long version will be higher than in the short version, cannot be supported. With regard to the different types of probe nonresponse, we find similar patterns across versions. The amount of complete nonresponse (respondents who skipped the probing question) was between 2% and 22% in the short and between 3% and 23% in the long version. Between 1% and 23% of respondents in the short version, and between 3% and 17% in the long version, refused to answer. Furthermore, "I don't know" answers per probe varied, with between 1% and 3% in the short and between 1% and 5% in the long version.

Number of uninterpretable answers. Besides not answering at all, there were also respondents who provided a response that was incomplete or uninterpretable in the context of the probing question (e.g., "Because I feel like that," "fitted best," or "dark figures"). Contrary to Hypothesis 1, the proportion of answers that were uninterpretable did not differ significantly between the two conditions, with an average of 11.4% in the short and an average of 10.6% in the long version ($\chi^2 = .278$,

Table 5. Number of Uninterpretable Answers Per Probe.

Probe Position	1/1	2/2	3/3	4/6	5/7	6/8	7/11	8/12	9/14	10/15	11/17	12/19	13/21
Type	G	C	G	G	G	G	S	G	G	C	G	G	G
Short	1.9	8.7	6.7	13.3	4.0	9.1	25.0	9.4	26.0	2.0	11.2	18.5	9.7
Long	2.8	9.8	10.6	12.5	7.3	5.1	23.5	5.3	12.5	5.3	15.2	8.6	9.2

Note. G = general probe; S = specific probe; C = comprehension probe.

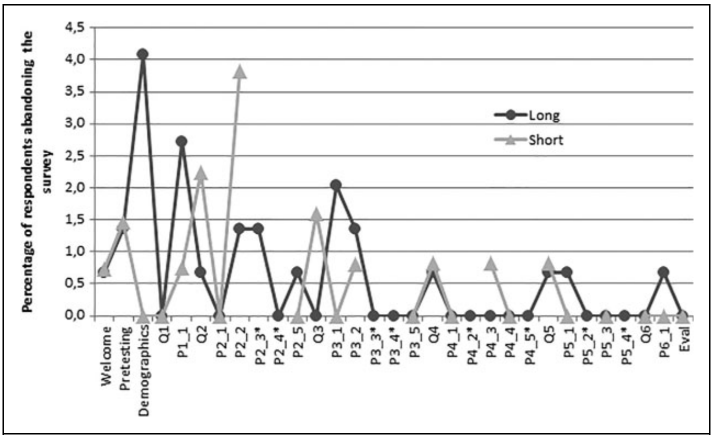

Figure 2. Dropouts per survey page. Note: Pages marked with * were only displayed in the long version.

$p = .598$, ranging from 1.9% to 26.0% in the short and from 2.8% to 23.5% in the long version, respectively). Table 5 shows the proportion of uninterpretable answers per probe. Using Bonferroni α correction, the proportion of uninterpretable answers did not differ significantly between the two conditions for any of the individual probing questions.

How many uninterpretable answers respondents provide does not seem to depend on the number of previous open-ended probing questions they have received, as the proportions are distributed very differently throughout the questionnaire. Instead, it seems to depend more on the nature of the probing question itself. By far, the largest number of noninterpretable statements occurred when answering the specific probe “Which public green areas did you have in mind when answering this question?” (Probe 7/11).

Dropouts. As mentioned above, we aimed for 120 completes in each condition. In total, 147 and 138 respondents started the survey in the long and in the short version, respectively. Thus, 27 respondents (18.4%) in the long and 18 respondents (13.0%) in the short version abandoned the questionnaire before completing it. However, in conflict with Hypothesis 1, this difference is not statistically significant ($\chi^2 = 1.517, p = .218$). The pattern of breakoff per survey page is presented in Figure 2. Breakoff does not seem to be a reaction to the number of preceding open-ended probing questions, as it does not increase during the survey. The highest dropout rates occur on the first pages of the survey in both versions. In the long version, after the screen with demographic questions, the highest number of dropouts occurred on the first page that had an open-ended probing question (Probe 1/1 on Question 1). In the short version, the highest number of dropouts occurred on the page that had the third open-ended probing question.

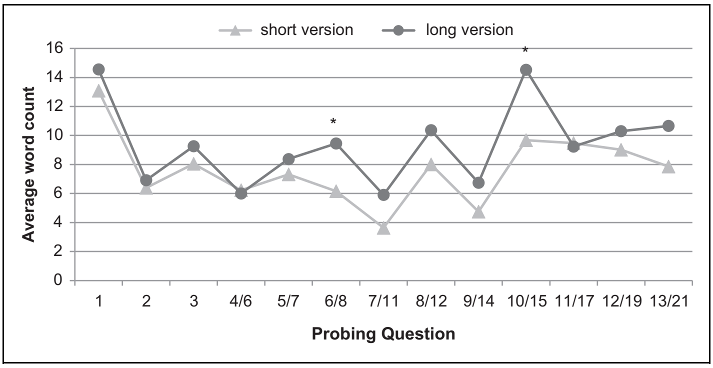

Figure 3. Average word count per probing question (short version=13 probes/long version=21 probes). Note: *Significant effect (p < .0038) after using Bonferroni-correction for multiple comparisons.

Word Count. Average word count was calculated by dividing the number of words respondents entered in response to the open-ended probes by the number of probes they had received. Meaningless answers such as "???,"" –," or "fgfsdg" and implicit refusals (respondents giving no answer whatsoever) received the value 0 on the word count variable. For most respondents, this meant dividing their overall response time by 13. However, some respondents skipped one or more questions and thus did not receive the corresponding follow-up probe. Also, respondents who answered that they do not use an automobile in Question 3 did not receive the following general probe. Hence, in some cases, response times were divided by a smaller number than 13.

Diverging from our first hypothesis (Hypothesis 1), respondents in the short condition did not enter more words into the open text boxes than respondents in the long condition ($t = -1.344$, $p = .180$). On average, respondents in the long condition produced 8.6 words per probe, while those in the short condition produced 7.5 words. With the exception of two probes (Probe 4/6 and Probe 11/17), respondents in the long condition entered more words into the open text boxes than respondents in the short condition (see Figure 3). Applying Bonferroni α correction, these differences on the item level are statistically significant for Probe 6/8 ($t = -3.36, p = .001$) and Probe 10/11 ($t = -3.32, p = .001$)

Response times. Response times were collected using Universal Client Side Paradata (Kaczmirek, 2005). In our analyses, we used respondents' mean response time, which was calculated by totaling the individual response times and dividing the overall response time by the number of probes respondents had received. Additionally, the upper and lower one percentiles of response times were defined as outliers (Lenzner, Kaczmirek, & Lenzner, 2010; Ratcliff, 1993) and excluded in the analyses. Hence, in some cases, response times were divided by a smaller number than 13.

Again contradicting our first hypothesis (Hypothesis 1), response times did not differ in the short and in the long condition ($t = -1.20, p = .231$). On average, respondents in the long condition spent 42.6 s per probe while those in the short condition spent 37.7 s per probe. Considering the individual probes, response times in the long condition were always longer than in the short condition, with the exception of probe 5/7 and probe 11/17. On the item level, no differences were statistically significant (see Figure 4).

Survey evaluation. Finally, we compared how respondents evaluated the survey overall. The question asked was "How do you rate the questionnaire overall?" with a 5-point scale ranging from *very poor*

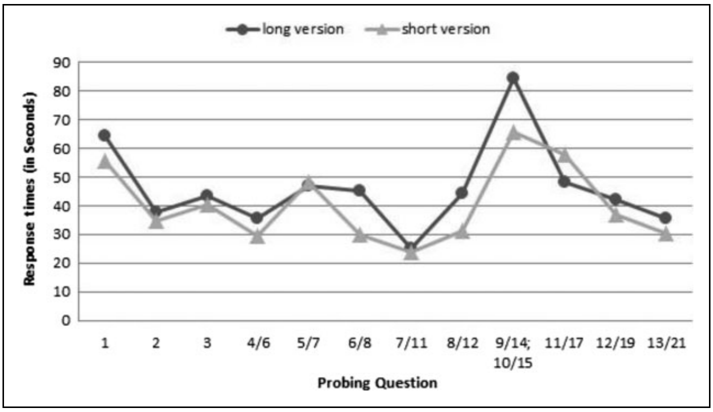

Figure 4. Average response times per probing question (short version=13 probes/long version=21 probes). Note: Probes 9 and 10 in the short and 14 and 15 in the long version were displayed together on one survey page.

to *very good*. Respondents in both groups evaluated the survey as equally “good” (3.85 in the short version vs. 3.93 in the long version; $F(1,239) = -.693$, $p = .245$). Thus, the number of open-ended probing questions did not affect overall satisfaction with the survey.

Content Analysis

In addition to the quantitative quality indicators, we examined whether the number of open-ended probing questions has an effect on the content or depth of respondents’ answers. The analysis focuses on differences in the topics respondents think of when they answer a survey question. For each of the 13 open-ended probing questions that were asked in both conditions,[2] a separate coding scheme was developed. The development of the code schemes and the analysis were guided by key questions such as “How do respondents justify/explain their answers?” or “What do respondents think of when answering?” The analysis is restricted to respondents who provided a response and were not classified as nonrespondents. The items were coded by one student assistant, and 33% of the data were coded a second time by the first author. The intercoder agreement varied between 85% and 100% for the 13 probes (short ∅ 95% vs. long ∅ 93.5%). To evaluate the differences between the two conditions, we conducted χ^2 tests for each probe. Overall, we found largely overlapping results in both questionnaire versions and we did not find significant differences in topics. Hypothesis 2, stating that the number of open-ended questions has an effect on the content of respondents’ answers, can thus not be supported.

We report the results of two probing questions (one positioned in the middle of the questionnaire and one at the end) as examples; the remaining results can be found in the Online Appendix. The results of the χ^2 tests are reported under each table. The first example is a specific probe that asked *What cases of fraud did you have in mind while answering this question*? as a follow-up to the question *What is your estimate of the probability that in the next 12 months you will become a victim of fraud (e.g., grandson scam, without cybercrime).* The probe was the 9th probing question in the short version and the 14th in the long version. Respondents in the short and in the long version had largely overlapping results (see Table 6). They mostly mention forms of fraud such as phishing or Internet fraud, confidence tricks, the grandparent scam, dubious telephone calls, and generally being cheated. A few respondents in both conditions thought of robbery and receiving too little change at the cash register.

Table 6. Probe P4_3 (9/14): Probability of Being a Victim of Deception in the Next 12 Months. Cases of Fraud (%).

Theme	Short ($N = 96$)	Long ($N = 88$)
Phishing/Internet fraud	31.3% (30)	21.6% (19)
Confidence trick	12.5% (12)	25.0% (22)
Grandparent scam	11.5% (11)	15.9% (14)
Dubious telephone calls	8.3% (8)	12.5% (11)
General: Cheated	7.5% (9)	6.7% (8)
Robbery	2.1% (2)	2.3% (2)
Attack	1.0% (1)	2.3% (2)
Change at the cash register	1.0% (1)	1.1% (1)
Useless/not codable	26.0% (25)	12.5% (22)

Note. $\chi^2 = 11.696$, $df = 89$, $p = 165$.

Table 7. Probe P6_1 (13/21): Satisfaction With Measures Implemented by Local Government for Achieving and Maintaining Public Security (%).

Theme	Short ($N = 93$)	Long ($N = 87$)
Unsatisfied because of concrete problems	32.3% (30)	26.4% (23)
Could be better/some good, some bad (no specification)	11.8% (11)	17.2% (15)
So far everything is fine (vague, sweeping overall judgment)	15.1% (14)	18.4% (16)
Not concerned with this issue, hence don't know	11.8% (11)	6.9% (6)
Concrete positive aspects are mentioned	7.5% (7)	16.1% (14)
Government is forced to satisfy all/different groups	7.5% (7)	2.3% (2)
Partly satisfied because of concrete problems	4.3% (4)	3.4% (3)
Useless/not codable	9.7% (9)	9.2% (8)

Note. $\chi^2 = 8.266$, $df = 79$, $p = .310$.

The second example reported is the last probe asked in each version, that is, the 13th probe in the short version and the 21st in the long version, administered after Survey Question 6. Survey Question 6 asked about respondents' satisfaction with measures implemented by their local government for achieving and maintaining public security. The probing question was a general probe, and respondents were asked to provide some information on why they had selected the answer category they had chosen, which ranged from *very satisfied* to *very dissatisfied*. As can be seen in Table 7, respondents in both conditions gave similar explanations. Most respondents stated that they were unsatisfied and named concrete problems; some reported that some things were good and some were bad; others said they had no reason to complain (Code *So far everything is fine*). There was no category that was mentioned only in one condition. Once again, the proportion of uncodable answers was similar between conditions.

Discussion

This study provides evidence about the willingness of respondents to answer a large number of open-ended probing questions in a cognitive online pretest and on how the number of probes affects the quality of the pretesting results. Contrary to our assumptions, whether there are 13 (short condition) or 21 open-ended probes (long condition) in a 26-item questionnaire has no effect on the quality of respondents' answers. Respondents in the long condition neither entered fewer words into their response fields nor did they spend less time answering the probes. In addition, neither the probe nonresponse rate nor the rate of uninterpretable answers differed between the two conditions. Only the dropout rate was slightly higher in the long than in the short condition (though this finding was not statistically significant). Therefore, for obtaining the intended net sample size in a cognitive online pretest, it might be necessary to oversample respondents, and to a greater extent, the more open-ended probes are being asked.

Also contrary to our assumptions, we found no differences either in the depth of respondents' answers to the probes (as measured by number of themes covered) nor in their overall satisfaction with the survey. All in all, these results suggest that respondents are willing to answer a relatively high number of open-ended probing questions and that asking (at least) up to 21 probes in a cognitive online pretest does not undermine the quality of respondents' answers to the probes. One limitation of this study is that survey questions and the corresponding probing questions were not randomized within versions. For this reason, it cannot be ruled out that respondents answered the open-ended probes the way they did because of the topic of the questions and not solely on the basis of the position in the questionnaire. Further studies should try to disentangle the effects of the position, type, and topic of the probing questions. Another possible limitation is the use of an online panel for recruiting respondents for the pretest: First, depending on the access panel policy, further invitations to other surveys conducted within the panel, or whether respondents receive an incentive or payment, may depend on the quality of survey participation. Second, online panelists might be comparably experienced in answering web surveys and their response behavior might differ from the general population. On the other hand, it is generally desirable in pretests (irrespective of whether they are conducted online or face-to-face) to recruit participants who are trained in answering survey questions (and cognitive probes) in order to get the most information out of each interview or completed questionnaire (cf. Willis, 2004). Hence, relying on online panelists instead of less experienced survey respondents may actually increase the validity of pretesting results. Again, this issue certainly demands future research. And finally, the increasing use of smartphones to fill out web surveys calls for research as to whether the devices used (e.g., PC, tablet, smartphone) have an effect on the quality of probe responses.

The present study adds to the existing literature showing that web probing is a valuable method for determining the reasons why respondents answer survey questions in the way they do. However, in addition to the issues mentioned above, there are still several research areas worth exploring in future studies. First, we still do not know how long a cognitive online pretest should *maximally* take (i.e., how many open-ended probes online respondents are maximally able and willing to answer). Second, with regard to the implementation of probes, it remains unclear how many probes should ideally be shown per survey page (i.e., one, two, three, or even more?). And finally, given that some respondents provide more informative responses to probes than others, it might be worthwhile to examine what respondent characteristics (e.g., sociodemographics, personality traits) are associated with high-quality answers to probes in cognitive online pretests. With this knowledge, participants for cognitive online pretests could be recruited in an efficient way.

Data Availability

The quantitative data set of this study is available on request from the corresponding author (Cornelia E. Neuert). The answers to the open-ended questions are not publicly available due to them containing information that could compromise participant privacy.

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: Part of this research was funded by the German Society for Online Research (DGOF).

Supplemental Material

Supplemental material for this article is available online.

Software information

All analyses in the present study were conducted using IBM SPSS Statistics Version 24.0.

Notes

1. Also called "online probing."
2. Of the 21 probes asked in the long condition, 13 were asked in the short condition and thus could be compared between the two conditions.

References

Behr, D., Bandilla, W., Kaczmirek, L., & Braun, M. (2014). Cognitive probes in web surveys: On the effect of different text box size and probing exposure on response quality. *Social Science Computer Review*, *32*, 524–533.

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.

Borg, I., & Zuell, C. (2012). Write-in comments in employee surveys. *International Journal of Manpower*, *33*, 206–220.

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.

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

Kaczmirek, L. (2005). A framework for the collection of universal client side paradata (UCSP). Retrieved March 11, 2019, from http://kaczmirek.de/ucsp/ucsp.html

Kwak, N., & Radler, B. (2002). A comparison between mail and web surveys: Response pattern, respondent profile, and data quality. *Journal of Official Statistics*, *18*, 257–273.

Lenzner, T., Disch, K., Gebhardt, S., & Menold, N. (2015). *AUDITS—Methodological tools for the definition of local security policies* . (Cognitive online pretest. GESIS Project Report [Version 1.0]). doi:10.17173/pretest12

Lenzner, T., Kaczmirek, L., & Lenzner, A. (2010). Cognitive burden of survey questions and response times: A psycholinguistic experiment. *Applied Cognitive Psychology*, *24*, 1003–1020.

Lenzner, T., & Neuert, C. E. (2017). Pretesting survey questions via web probing—Does it produce similar results to face-to-face cognitive interviewing? *Survey Practice*, *10*, 1–11.

Meitinger, K., & Behr, D. (2016). Comparing cognitive interviewing and online probing: Do they find similar results? *Field Methods*, *28*, 363–380.

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* (pp. 215–244). New York, NY: Routledge.

Ratcliff, R. (1993). Methods for dealing with reaction time outliers. *Psychological Bulletin*, *114*, 510–532.

Reja, U., Manfreda, K. L., Hlebec, V., & Vehovar, V. (2003). Open-ended vs. close-ended questions in web questionnaires. *Metodoloski zvezki*, *19*, 159–177

Scholz, E., & Zuell, C. (2012). Item non-response in open-ended questions: Who does not answer on the meaning of left and right? *Social Science Research*, *41*, 1415–1428.

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

Willis, G. B. (2004). *Cognitive interviewing: A tool for improving questionnaire design*. Sage Publications.

Züll, C. (2016). *Open-ended questions. GESIS survey guidelines*. Mannheim, Germany: GESIS—Leibniz Institute for the Social Sciences. doi:10.15465/gesis-sg_en_002

Züll, C., Menold, N., & Körber, S. (2014). The influence of answer box size on item nonresponse to open-ended questions in a web survey. *Social Science Computer Review*, *33*, 115–122.

Author Biographies

Cornelia E. Neuert is a senior researcher and team leader at GESIS—Leibniz Institute for the Social Sciences. Her research interests focus on questionnaire design and evaluation, in particular cognitive testing, eye tracking, and web probing. She can be contacted at cornelia.neuert@gesis.org

Timo Lenzner is a senior researcher at the GESIS Pretest Lab, GESIS—Leibniz Institute for the Social Sciences. His research interests focus on questionnaire design and evaluation, web surveys, and usability. He can be contacted at timo.lenzner@gesis.org