

Open-ended versus Closed Probes: Assessing Different Formats of Web Probing

Sociological Methods & Research

2023, Vol. 52(4) 1981–2015

© The Author(s) 2021



Article reuse guidelines:

sagepub.com/journals-permissions

DOI: 10.1177/00491241211031271

journals.sagepub.com/home/smr

Cornelia E. Neuert¹ , Katharina Meitinger² 
and Dorothee Behr¹

Abstract

The method of web probing integrates cognitive interviewing techniques into web surveys and is increasingly used to evaluate survey questions. In a usual web probing scenario, probes are administered immediately after the question to be tested (concurrent probing), typically as open-ended questions. A second possibility of administering probes is in a closed format, whereby the response categories for the closed probes are developed during previously conducted qualitative cognitive interviews. Using closed probes has several benefits, such as reduced costs and time efficiency, because this method does not require manual coding of open-ended responses. In this article, we investigate whether the insights gained into item functioning when implementing closed probes are comparable to the insights gained when asking open-ended probes and whether closed probes are equally suitable to capture the cognitive processes for which traditionally open-ended probes are intended. The findings reveal statistically significant differences with

¹ GESIS—Leibniz Institute for the Social Sciences, Mannheim, P.O. Box 12 21 55, 68072 Mannheim, Germany

² Utrecht University, the Netherlands

Corresponding Author:

Cornelia E. Neuert, GESIS—Leibniz Institute for the Social Sciences, Mannheim, P.O. Box 12 21 55, 68072 Mannheim, Germany.

Email: cornelia.neuert@gesis.org

regard to the variety of themes, the patterns of interpretation, the number of themes per respondent, and nonresponse. No differences in number of themes across formats by sex and educational level were found.

Keywords

web probing, open-ended questions, embedded closed probes, cognitive interviewing, web survey

Cognitive pretesting is a valuable tool to assure data quality of survey instruments prior to data collection (Groves et al. 2011). The method of web probing is a recent addition to the pretesting toolbox of questionnaire designers. Web probing is defined as implementing probes, which are typically used in cognitive interviewing, in web surveys (Behr et al. 2017). In traditional “in-person” cognitive interviewing, probes, that is, open-ended follow-up questions, are administered by an interviewer “to collect additional verbal information about the survey responses, which is used to evaluate the quality of the response or to help determine whether the question is generating the information that its author intends” (Beatty and Willis 2007:288). Hence, the main goal of asking probes is to evaluate the comprehensibility and validity of a given survey item. Now a widespread pretesting technique, the use of probes to better understand respondents’ survey answers dates back to Schuman (1966, in his case “random probes”) and Converse and Presser (1986). Recently, the method of web probing, which has the same goals than cognitive interviewing as described above, has risen to popularity due to several benefits compared to traditional “in-person” cognitive interviewing, such as a facilitated recruitment of respondents via an online survey provider and a relatively fast data collection. It also allows for broader geographical coverage and larger sample sizes, which makes an assessment of the prevalence of themes and a detailed analysis of subgroups and answer patterns possible (Behr et al. 2017; Edgar, Murphy, and Keating 2016; Lee et al. 2020; Meitinger and Behr 2016; Meitinger, Braun, and Behr 2018). On the downside, open-ended questions in web surveys, such as probes, pose additional burden on respondents in comparison to answering closed questions. This is indicated by increased levels of nonresponse or by responses that do not correspond to the probe type (mismatching response) due to the lack of an interviewer who could motivate respondents and clarify the meaning of probes if necessary (Behr et al. 2014; Lenzner and Neuert 2017; Meitinger and Behr 2016). In contrast,

nonresponse in “in-person” cognitive interviewing studies occurs rarely (Lenzner and Neuert 2017; Meitinger and Behr 2016). Web probing also requires cleaning and coding of open-ended responses, making the qualitative analysis of these large sample sizes time- and labor-intensive (Meitinger 2017).

Against these backdrops, Scanlon (2019, 2020a) proposed a novel approach based on implementing web probing with so-called targeted embedded probes, that is, with closed instead of open-ended probes to understand how respondents answer survey items. The closed probes in Scanlon’s studies are developed based on previous “in-person” cognitive interviewing results. That is, respondents’ perspectives that are revealed during traditional cognitive interviewing serve as the response categories of the targeted embedded probes.¹ According to Scanlon (2019), embedding these closed probes alongside the questions to be tested allows quantifying the patterns of interpretation that respondents use as well as the occurrence of errors in a larger survey population (Scanlon 2019). Statistical analyses of this kind are not possible with the rather small samples of 5–30 respondents (Willis 2005) that are typically used for cognitive interviewing studies. For those carrying out a pretesting study, embedding closed probes drastically reduces the costs and burden involved in the data processing and analysis stages of open-ended probes because this method does not require manual coding of open-ended responses. Closed probes also help to prevent coding challenges such as lacking intercoder reliability or noninterpretable responses. Additionally, from a respondents’ perspective, the response burden is reduced when closed compared to open-ended questions need to be answered (Bradburn 1978).

However, a research gap exists as to whether the insights gained into item functioning when implementing closed probes are comparable to the insights gained when asking open-ended probes. This question is important to address if embedded closed probes are to supplement the pretesting toolbox of researchers who have hitherto focused on open-ended probing to understand the underlying cognitive processes by not restricting or guiding the respondents’ answers in any way.

Probes as a Special Type of Open-ended Questions in the Context of Web Probing

Due to the intention of collecting nondirective information, that is, information that does not influence respondents’ answers, probes are commonly administered as open-ended questions (see Willis 2005:48 for probe examples). According to Tourangeau’s four-stage question–response process model, respondents have to complete four distinct steps in order to answer

a survey question: Respondents must comprehend the question, retrieve relevant information, make use of the information to form a judgment, and answer the question by selecting a response (Tourangeau 1984). Open-ended probes seek to obtain additional information on a question under evaluation, in particular, related to potential problems during the above-mentioned question–response process. Probes are often designed to investigate one particular cognitive process (e.g., there are comprehension probes such as “What do you understand by term X?” or recall probes such as “How do you remember this?”; Willis 2005). In contrast to standard open-ended survey questions that aim, for instance, to measure knowledge, unaffected opinions, or an unknown range of possible answers (Geer 1991; Reja et al. 2003; Züll 2016), probes aim to gather information on respondents’ thought processes when answering a survey question and to compare whether the survey question is understood as intended. Probes can be administered concurrently (i.e., the probe follows immediately after a respondent has selected a response from a closed-ended survey item) or retrospectively (i.e., after the respondents has answered the entire questionnaire). While retrospective probing is intended not to disrupt the flow of the whole questionnaire, concurrent probing is meant to ensure that the thought process is still available in short-term memory (Willis 2005). All probing varieties follow the assumption that the answers given in response to a probe are indeed related to the answers respondents provided to a preceding survey question (see Silber, Züll, and Kühnel 2020 for a recent demonstration using voting behavior).

The use of probes in online surveys differs in several aspects from the traditional “in-person” cognitive interviewing approach. In the context of “in-person” cognitive interviews, interviewers can clarify respondents’ ambiguous answers; follow up on unexpected, short, or unclear statements; and motivate respondents to finish the cognitive interview. In the context of web probing, however, these spontaneous follow-ups are not possible, and hence, nonresponse rates as well as the number of uninterpretable answers are often more elevated than in cognitive interviews (Lenzner and Neuert 2017; Meitinger and Behr 2016). Answering probes in a self-administered format requires respondents to provide their response by writing/typing it into an open-ended text box instead of responding in a more conversational form.² Also, the method is less flexible than “in-person” interviews as web probes need to be developed and programmed in advance (they are “anticipated probes” in contrast to “spontaneous” or “emergent probes” that depend on respondents behavior and are not scripted prior to the interview; see Willis, 2005, or Beatty and Willis 2007).

From the literature on open-ended survey questions! in general, it is known that formulating answers in one's own words poses a higher burden on respondents' cognitive abilities than selecting a response category from a set of provided options (Bradburn 1978), and it also requires both more time and more consideration (Holland and Christian 2009). Respondents also cannot use a provided list of answer options to infer the meaning of the question (Smyth et al. 2009). Likewise, the answer options cannot serve as a reminder for themes that respondents might not have considered otherwise (Schwarz 1999). The higher perceived difficulty and cognitive effort required when answering open-ended questions lead to a lower willingness to respond (Dillman, Smyth, and Christian 2009; Galesic 2006; Züll 2016), which is reflected in higher rates of item nonresponse (Andrews 2005; Borg and Züll, 2012; Denscombe 2008; Reja et al. 2003; Scholz and Züll 2012) and survey break-offs (Crawford, Couper, and Lamias 2001; Knapp and Heidingsfelder 2001) compared to closed questions. Answering cognitive probes may even be more burdensome for respondents than standard open-ended survey questions as reporting on response processes might be especially taxing. Furthermore, the perceived difficulty of providing responses in one's own words, and hence, the quality of the response to open-ended questions can be affected by sociodemographic characteristics such as sex, age, and education (Denscombe 2008; Stern, Dillman, and Smyth 2007). For "standard" open-ended survey questions, research outcomes are mixed: Some studies suggest that there is no sex difference regarding the likelihood of answering open-ended questions. However, in terms of length, some studies found women to provide longer responses than men (Denscombe 2008; Stern et al. 2007), while others did not find such differences (Oudejans and Christian 2010). With regard to education (or underlying cognitive abilities), previous findings indicate that respondents with higher educational qualifications provide longer responses and less item nonresponse (Scholz and Züll 2012; Stern et al. 2007). In the context of web probing, previous findings are also mixed regarding nonresponse and length of answers provided in the text fields. Comparing data sets of six pretests revealed that female respondents, older respondents, and higher educated respondents sometimes wrote longer responses into the text boxes (Lenzner and Neuert 2019), while higher educated as well as older respondents were less likely to leave the open-ended probes blank (Lenzner and Neuert 2019; Meitinger and Kaczmirek 2018).

Given the challenges or limitations inherent in "standard" open-ended questions and cognitive probes, it is not surprising that closed probes are looked at as an alternative. Using closed probes as a substitute for open-ended probes presumes that the initial in-person cognitive interviews illicit

themes similar to those that respondents would mention when asked open-ended probes. The usefulness of closed probes also depends on the researchers' skills to verbalize these themes in appropriate categories that respondents would bring up and, hence, the same cognitive processes, as when asking open-ended probes. To the best of our knowledge, up to now, the open-ended and the closed web probing approach have not yet been systematically compared, which is the goal of this study.

Research Questions and Hypotheses

This study aims to assess the use of targeted embedded closed probes, compared to open-ended probes. We will look into the comparability of the responses and what the findings mean for the practice of online pretesting.

Research Question 1: In web probing studies, will open-ended probes and closed probes capture a comparable number and type of substantive themes?

Research Question 2: In web probing studies, are open-ended probes and closed probes comparable with regard to nonresponse rates?

Furthermore, we will specifically analyze the subgroups of lower educated and older respondents to assess whether the method of closed probes is cognitively less demanding for those groups than open-ended probes.

Based on the findings from previous studies that compared standard open-ended and closed survey questions and based on the assumptions which are made for targeted embedded probes, we put forward the following hypotheses in the next section.

Coverage of Themes (New Themes and Subcategories)

The method of closed probes assumes theoretical topic saturation due to the fact that the closed set of response options has been developed based on previously conducted in-person cognitive interviews. However, topic saturation highly depends on the sample composition and the number of cognitive interviews conducted (Guest, Bunce, and Johnson 2006). Furthermore, previous research comparing open-ended and closed question formats has shown that the answers to the open-ended format resulted in additional categories not covered by the closed format (Reja et al. 2003; Schuman and Presser 1979). Moreover, by allowing respondents to elaborate on their responses, open-ended questions can provide more detailed information from

a respondent on a topic of interest resulting in more diverse (sub-)themes as well as broader information (Friborg and Rosenvinge 2013; Schmidt, Gummer, and Roßmann 2020).

The open-ended probe format allows examining which themes respondents were thinking of without being influenced by the response categories provided (Geer 1991). Offering response suggestions might lead respondents to select themes that they did not think of when answering the closed question in the first place. This would also affect the number of themes mentioned (see Hypothesis 4 in Number of Themes Mentioned per Respondent subsection).

We expect the open-ended probe format to generate additional categories that cannot be subsumed under the response categories provided in the closed format, taking into account that the closed format is limited in the number of categories presented. We also expect respondents in the open-ended condition to mention additional sub-themes by being more specific in their responses. We anticipate that this specificity will result in subcategories that were not provided in the initial set of categories in the closed format. This leads to the following hypothesis regarding the variety of themes with two subhypotheses.

Hypothesis 1: In web probing studies, responses to open-ended probes will cover a wider variety of themes than those provided in the closed format.

Hypothesis 1a: In web probing studies, responses to open-ended probes will capture themes that are distinct from response categories in the closed format.

Hypothesis 1b: In web probing studies, responses to open-ended probes will allow to identify more specific themes than responses to closed probes, leading to additional and more detailed subcategories of the main categories in the open-ended format.

Patterns of Interpretation (Response Distribution and Ranking of Themes)

Previous research on closed versus open-ended survey questions has shown that frequency distributions are often not comparable across these two formats. Asking people to name “the most important problems facing the country” in a split-ballot experiment led, for example, to 22 percent affirmatively selected responses for shortages in food and energy, while this

response was mentioned by only 0.2 percent of respondents when no set of response categories was provided (Schuman and Presser 1979). A second study by Schuman and Presser (1981) revealed that almost 60 percent of responses that were given in an open format could not be coded into the five answer categories provided in the closed format. Similar differences in response distributions were later also found for comparisons of open-ended and closed formats in web surveys (Reja et al. 2003).

While these findings pertain to “standard” open-ended survey questions, we have to consider in our study that the cognitive process in responding to a probe is slightly different than the one in responding to an open-ended survey question. For the probe, respondents are supposed to relate their response back to the previously answered survey question. Therefore, the variety of themes might be narrower than for a “standard” open-ended survey question. This potentially renders the underlying associations and, hence, the response distributions more comparable. The results should be consistent, at least in terms of what is most often mentioned in relative terms (ranking of themes).

Hypothesis 2: In web probing studies, the response distributions will be different across open-ended and closed probe formats.

Hypothesis 3: In web probing studies, the most often mentioned categories will be identical in both probe formats.

If the comparisons between both probe formats reveal that the response distributions are not comparable, this might be an indication that topic saturation was not reached during the cognitive interviewing stage or that different cognitive processes take place across these formats.

Number of Themes Mentioned per Respondent

Regarding the number of themes mentioned by individual respondents, we expect, based on previous research, the following two mutually reinforcing processes. By providing response categories, respondents are reminded of aspects that they would not have thought of otherwise (Schwarz 1999). And, due to the higher effort to type/write an answer manually into the textbox compared to simply having to report a theme by selecting the respective option (e.g., Dillman et al. 2009; Reja et al. 2003), we expect that respondents will mention fewer themes in the open-ended format than in the closed format.

Hypothesis 4: Respondents answering the closed probes will select more response options than respondents answering the open-ended probes will mention themes.

Nonresponse

Open-ended survey questions in general as well as probes in particular suffer from higher rates of nonresponse due to higher response burden (e.g., Denscombe 2008; Meitinger and Behr 2016; Neuert and Lenzner 2019; Reja et al. 2003). We expect a similar effect for open-ended versus closed probes.

Hypothesis 5: Open-ended probes will be more affected by probe nonresponse than closed probes.

Impact of Sociodemographic Characteristics (Themes and Nonresponse)

Finally, we examine whether there is an interaction effect between socio-demographic characteristics and probe format on both the number of themes mentioned and the amount of nonresponse observed. Based on previous findings regarding the effect of sociodemographic characteristics on the quality of the response to open-ended questions in the context of web probing, we expect that respondents with lower educational qualifications provide shorter responses, and hence fewer themes, and higher item nonresponse in the open-ended compared to the closed format, while we do not expect differences across formats for respondents with higher educational qualifications. Due to the mixed findings regarding sex and age, we do not postulate hypotheses, even though we will study the impact of these variables, too.

Hypothesis 6: We expect the closed probes to be easier for lower educated respondents resulting in less nonresponse and more themes than for the open-ended probes.

Method

The study was conducted with respondents from the German opt-in online panel of the Respondi AG, which is a commercial ISO-certified panel provider (respondi.com). The questionnaire contained questions on the topic of general health and was fielded between July 1 and July 9, 2019. The web survey used quotas for sex, age, and education. Overall, 2,183 panelists accepted the survey invitation, of whom 186 were screened out, 237 broke

off, and 1,760 completed the survey. The break-off rate was 11.9 percent (cf. Callegaro and DiSogra 2008). The mean duration of the questionnaire completion was 8 minutes and 27 seconds.

Measurement Instruments

For the comparison of probe formats, we implemented three fully randomized, between-subject experiments. We selected three survey questions for which there were already closed probes available. The closed probes had been developed and proposed by Scanlon and colleagues (2020a, 2020b) for the U.S. context.³ We translated these into German, following the double translation and team review approach (Harkness 2003) involving the three authors of this study, and tested them for cultural transferability (i.e., Are the response categories understood as intended?) in two rounds of cognitive interviews in the GESIS pretest lab in Germany (see Section B of Appendix for further details, which can be found at <http://smr.sagepub.com/supplemental/>). The survey questions used in the experiments were questions asking for (1) self-rated health, (2) self-rated pain, and (3) self-rated physical activity (two separate questions asking for light/moderate and vigorous activities; see Table 1 for the question and probe wording). Probes were administered concurrently.

Experiment 1. Self-rated health is one of the most frequently used health measures (Garbarski et al. 2017) and is asked in many cross-national surveys (e.g., SHARE, the Survey of Health, Ageing and Retirement in Europe, ISSP, the International Social Survey Program, or ESS, the European Social Survey). The question on self-rated health was as follows: “Would you say your health in general is excellent, very good, good, fair, or poor?”

The *closed* follow-up probe was as follows: “When you answered the previous question about your health, what did you think of?” A set of the following nine closed answer categories was provided:

1. Diet and nutrition
2. Exercise habits
3. Smoking or drinking habits
4. Health problems or conditions
5. Lack of health problems or conditions
6. Pain: The amount of pain that you have
7. Ability to do activities daily living (ADL) without assistance
8. Sleep: The amount of sleep you get
9. Mental or emotional health

Table 1. Question Wording of Survey Questions and Probes.

	Experiment 1	Experiment 2	Experiment 3
Target question	Would you say your health in general is very good, good, fair, poor, or very poor?	In the past [3 or 6] months, how often did you have pain?	How often do you do light or moderate leisure time physical activities for at least 10 minutes that cause only light sweating or a slight to moderate increase in breathing or heart rate?
Response options	very good – good – fair – poor – very poor	never – some days – most days – every day	[Number box] Number of times never – per day – per week – per month – per year – unable to do this activity [Dropdown list]
Closed probe			
Question stem	When you answered the previous question about your health, what did you think of?	Which of the following statements, if any, describe your pain in the past 3/6 months?	Which of the following types of physical activity, if any, did you include when you answered the previous question?
Response options	<ol style="list-style-type: none"> 1. Diet and nutrition 2. Exercise habits 3. Smoking or drinking habits 4. Health problems or conditions 5. Lack of health problems or conditions 6. Pain: The amount of pain that you have 7. Ability to do daily living without assistance 8. Sleep: The amount of sleep you get. 	<ol style="list-style-type: none"> 1. It is constantly present 2. Sometimes I'm in a lot of pain and sometimes it's not so bad 3. Sometimes it's unbearable and excruciating 4. When I get my mind on other things, I'm not aware of the pain 5. It is occasional and does not last completely 6. Medication can take my pain away 7. My pain is because of my current or past work 8. My pain is because of exercise 	<ol style="list-style-type: none"> 1. Running or jogging 2. Hiking 3. Walking as part of your job 4. Walking outside of work 5. Yardwork or cleaning your house 6. Working out with exercise equipment 7. Lifting weights 8. Cycling, swimming, or other aerobic activities

(continued)

Table 1. (continued)

	Experiment 1	Experiment 2	Experiment 3
	9. Mental or emotional health	9. My pain was caused by a recent injury or infection 10. My pain is minor and infrequent	9. Yoga or stretching 10. Playing a sport, namely [text box] 11. Other [text box]
Open probe	When you answered the previous question about your health, what did you think of?	Could you describe your pain in more detail?	Which types of physical activity, did you think of when you answered the previous question?
Target question			How often do you do vigorous leisure-time physical activities for at least 10 minutes that cause heavy sweating or large increases in breathing or heart rate? [Number box] Number of times
Response options			never – per day – per week – per month – per year – unable to do this activity [Dropdown list]
Closed probe			Which of the following types of physical activity, if any, did you include when you answered the previous question?
Question stem			1. Running or jogging
Response options			2. to 11 identical to previous question
Open probe			Which types of physical activity, did you think of when you answered the previous question?

To allow respondents in the closed probe format to indicate whether they had thought of anything else beyond the provided categories—and without disrupting their response behavior or endangering comparability across probe formats—we implemented an additional open-ended question on the following screen asking whether respondents had thought of anything else (additional follow-up probe).

The wording of the *open-ended* probe question was identical to the question stem of the closed probe (“When you answered the previous question about your health, what did you think of?”). However, in the open-ended format, respondents were provided with a text box to enter their response. The text box had a width of 80 columns and a height of 5 rows, but no restriction in writing space, which could go beyond this size.

Experiment 2. The survey question asking for self-rated pain was a split-ballot experiment containing two different time frames and was worded as follows: “In the past 3/6 months, how often did you have pain?” Respondents in both experimental conditions—closed versus open-ended probe format—were randomly assigned to one version asking either for pain in the last six or last three months, resulting in four experimental groups. For the analyses of the differences between closed and open-ended probes, we do not distinguish between the two different time frames.

The *closed* probe following the question on self-rated pain was: “Which of the following statements, if any, describe your pain in the past 3/6 months?” The following ten statements were provided:

1. It is constantly present
2. Sometimes I’m in a lot of pain and sometimes it’s not so bad
3. Sometimes it’s unbearable and excruciating
4. When I get my mind on other things, I’m not aware of the pain
5. It is occasional and does not last
6. Medication can take my pain away completely
7. My pain is because of my current or past work
8. My pain is because of exercise
9. My pain was caused by a recent injury or infection
10. My pain is minor and infrequent

In case the respondents in the closed-ended format did not select one of the provided responses, they were asked an additional open-ended probe on the next survey page, asking how they would describe the pain in their own words (additional follow-up probe).

When it comes to the *open-ended* probe format, the following applied: The wording of the open-ended probe was as follows: “Could you describe your pain in more detail?” Even though this wording slightly differs from the closed probe wording, we tried to be as similar as possible to ensure comparability between probe formats. The probing questions—either closed or open-ended—were shown to those respondents who had answered that they had suffered from pain at least some days in the past three or six months.

Experiment 3. The questions on self-rated physical activities were asked for light/moderate and for vigorous activities. The survey question asking for light/moderate activities was as follows: “How often do you do light or moderate leisure time physical activities for at least 10 minutes that cause only light sweating or a slight to moderate increase in breathing or heart rate?” The survey question asking for vigorous activities was as follows: “How often do you do vigorous leisure-time physical activities for at least 10 minutes that cause heavy sweating or large increases in breathing or heart rate?”

Following each of the self-rated physical activity questions, a *closed* probe was asked: “Which of the following types of physical activity, if any, did you include when you answered the previous question?” The following set of activities was provided:

1. Running or jogging
2. Hiking
3. Walking as part of your job
4. Walking outside of work
5. Yardwork or cleaning your house
6. Working out with exercise equipment
7. Lifting weights
8. Cycling, swimming, or other aerobic activities
9. Yoga or stretching
10. Playing a sport (semi-open)
11. Other (semi-open)

For two categories (“playing a sport” and “other”), the format was semi-open and included a text box for respondents to specify which sport they had played and to be able to mention other activities if these were not part of the response categories provided. This information can be used as an additional indicator for the appropriateness of the closed response categories provided and whether respondents use it to provide more detailed responses.

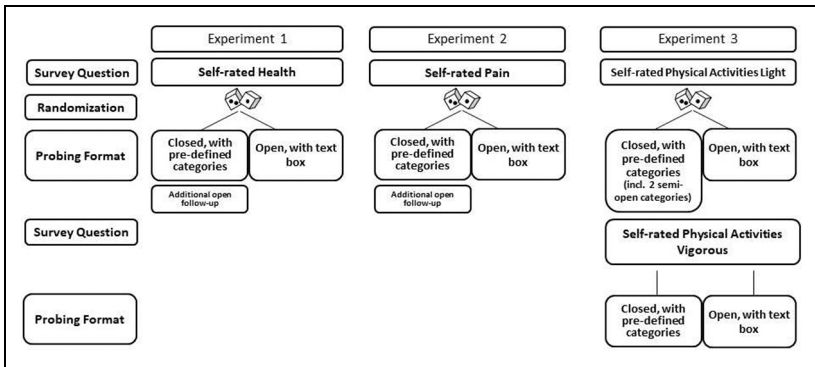


Figure 1. Experimental design for the comparison of open-ended versus closed probes.

The *open-ended* probes following each of the self-rated physical activity questions were as follows: “Which types of physical activity did you think of when you answered the previous question?” The overall design of the experiments is illustrated in Figure 1.

The closed probes in all experiments used a check-all-that-apply question format, and respondents could tick as many options as they wanted. Answering the probes—whether open-ended or closed—was voluntary. However, we implemented so-called soft prompts in our survey. The survey software automatically checked whether the respondent had answered a question. If this was not the case, a message was displayed stating “Please answer all questions.” Nevertheless, respondents could decide to skip the question without answering by checking a box indicating that they want to leave this question blank.

For each of the three experiments, respondents were randomly assigned to the different probe formats. Of 856 respondents administered the self-rated health item, 414 respondents received the closed format, 442 the open-ended form. Of 1,435 respondents who answered the question on pain, 706 received the closed probe and 729 the open-ended probe. Of 1,760 respondents who answered the two question versions (light/moderate and vigorous) on activities, 883 received the closed format and 877 the open-ended form.

Coding Procedure, Measures, and Analytical Strategies

For data coding of the open-ended probe answers, we developed separate coding schemes. The coding schemes consisted in the first place of the

response options from the closed probes to ensure comparable analysis of probe answers across the formats. Additionally, the schemes were extended with new categories or refined with subcategories where needed based on the answers to the open-ended probes. The coding scheme for the self-rated health question was based on the scheme developed by Lee et al. (2020) and adapted for our purposes. The other two schemes (pain/activities) were each developed by one of the authors of this study and reviewed by the other two authors. Besides substantive codes, each scheme contained codes for non-response (including nonsubstantive or not interpretable responses). For the questions on self-rated health and activities, the order of the themes mentioned was also coded. All answers to the open-ended probes were coded by one of the authors and double-coded by a student assistant. The intercoder agreement varied between 93 percent for self-rated health and 99 percent for self-rated pain; agreement for self-rated physical activity was 97 percent. Discrepancies were discussed to make a decision about the final codings.

We coded each answer provided in response to an open-ended probe. Respondents could mention multiple themes, which were then coded separately. Each code equals one theme. In the closed probe format, *number of themes* corresponds to the number of response categories selected. In the case of the additional follow-up probe, all further aspects mentioned were coded as separate themes, like in the open-ended probe format. *Coverage of themes* is measured by comparing whether respondents in the open-ended format mention *new* themes that are not covered by the response categories provided in the closed format. *Response distributions* of themes are compared by contrasting the frequency of each response category selected in the closed-ended probe format with the frequency of that same category in the open-ended format. For each probe format, the categories are then *ranked* to compare their relative frequencies. Response categories selected with the same relative frequency receive the same ranking number, and then a gap is left in the ranking numbers (to indicate the shared ranking). For all three experiments, the comparison of response distribution and ranking are restricted to those categories that were provided in the closed format. To measure the amount of probe *nonresponse*, we divided all answers in either substantive responses or nonresponse. Nonresponse is an exclusive category and included no (substantive) answer at all (e.g., empty answer boxes, “-;” “. . . .;” “no comment”), random characters (“dlfgfdg”), don’t know responses (“no idea”; “unsure”), refusals (“don’t want to answer”), and responses that were insufficient for substantive coding (e.g., “Ok, thanks,” “just like that”).

All statistical analyses were performed with Stata (Version 15.1). To compare response distributions and the amount of nonresponse, we report Pearson's chi-square tests. Since we have directed hypotheses, we report one-tailed *t*-tests for the comparison of number of themes. To analyze whether there is an interaction between sociodemographic characteristics and probe format, two-way analyses of variance (ANOVAs) are performed. Except for the analyses of nonresponse, analyses are restricted to respondents who provided a substantive response.

Results

Coverage of Themes (New Themes and Subcategories)

Hypothesis 1a, which postulates that responses to open-ended probes capture themes that are distinct from response categories in the closed format, and Hypothesis 1b, which states that responses to open-ended probes will allow to identify more specific themes than closed probes leading to additional subcategories in the open-ended format, are confirmed for all three experiments. The overview of response categories presented in Tables 2–5 shows that for all three experiments, new categories arose from coding the open-ended probe responses. In Experiment 1, the closed probe following the self-rated health question contained nine categories or patterns of interpretation. Coding of the responses provided in the open-ended format resulted in six new categories. Of the respondents in the closed format, at least 8.5 percent used the additional open-ended follow-up probe. This resulted in one additional category (“handicap”) that was, however, not mentioned in the open-ended probe.

In Experiment 2, all respondents who answered the closed probe following the self-rated pain question selected at least one of the 10 provided categories, and hence, none of those respondents received the additional open-ended follow-up probe. Therefore, the three new categories described in the following exclusively emerged in response to the open-ended probe. The three new patterns were *pain location*, *pain symptoms*, and *consequences of pain*. Although the closed response options and hence the coding scheme covered aspects of *causes of pain*, *pain intensity*, and *pain duration*, respondents named additional aspects that were not already covered by the given response options. For instance, respondents described their *pain duration*, such as having “frequent” or “sudden” pain, but these responses could not be coded into the three provided options from the closed probe that covered pain duration: “1. It is constantly present,” “5. It is occasional and does not last,”

Table 2. Percentage of Codes Selected/Mentioned (Patterns of Interpretation) on the Probe following the Self-reported Health Question.

Categories	Closed (N = 414)		Open following Closed Probe		Open (N = 442)		Difference	
	%	Rank	%	Rank	%	Rank		
1. Diet and nutrition	35.3	3	0.2	5	2.0	5	34.8	$\chi^2 = 159.18, p < .001, \text{Cramer's } V = .43$
Weight	—	—	0.5	—	4.3	—	—	—
2. Exercise	37.9	2	0.7	4	3.2	4	34.3	$\chi^2 = 161.53, p < .001, \text{Cramer's } V = .43$
In/out of shape	—	—	—	—	6.3	—	—	—
Activity level	—	—	0.2	—	—	—	—	—
3. Smoking or drinking habits	22.7	6*	—	7	0.9	7	22.5	$\chi^2 = 100.21, p < .001, \text{Cramer's } V = .34$
4. Health problems or conditions	50.0	1	1.5	1	26.7	1	26.4	$\chi^2 = 49.29, p < .001, \text{Cramer's } V = .24$
5. Lack of health problems or conditions	26.3	4	—	6	1.4	6	24.4	$\chi^2 = 144.62, p < .001, \text{Cramer's } V = .37$
6. Pain: The amount of pain that you have	22.7	6*	0.2	2	6.6	2	16.5	$\chi^2 = 45.28, p < .001, \text{Cramer's } V = .23$
7. ADL without assistance	7.0	9	—	9	0.2	9	6.8	$\chi^2 = 29.05, p < .001, \text{Cramer's } V = .18$
8. Sleep: The amount of sleep you get.	16.4	8	0.5	8	0.5	8	15.3	$\chi^2 = 72.63, p < .001, \text{Cramer's } V = .29$
9. Mental or emotional health	23.0	5	1.0	3	4.3	3	20.4	$\chi^2 = 64.40, p < .001, \text{Cramer's } V = .27$
New categories:								
Health status, general ("My general state of health," "I feel healthy")	—	—	0.7	—	23.1	—	—	—
Life situation	—	—	—	—	—	—	—	—
Societal/demographic reasons	—	—	1.2	—	3.4	—	—	—
Health service usage	—	—	0.5	—	2.9	—	—	—
Time-related factors	—	—	0.5	—	2.5	—	—	—
Insufficient knowledge about health	—	—	—	—	—	—	—	—
Handicap	—	—	0.7	—	—	—	—	—

Note: * = shared Rank #6. Subcategories are indented and values of subcategories are in italics.

Table 3. Percentage and Ranking of Codes Selected/Mentioned (Patterns of Interpretation) on the Probe following the Question Asking for Pain.

Categories	Closed (N = 706)		Open (N = 729)		Rank Difference	
	%	Rank	%	Rank		
1. It is constantly present	9.1	9	2.1	6	7.0	$\chi^2 = 33.86, p < .001$, Cramer's V = .15
2. Sometimes I'm in a lot of pain and sometimes it's not so bad	32.9	2	3.6	3	29.3	$\chi^2 = 208.72, p < .001$, Cramer's V = .38
3. Sometimes it's unbearable and excruciating	11.3	6	0.3	8	11.0	$\chi^2 = 81.39, p < .001$, Cramer's V = .24
4. When I get my mind on other things, I'm not aware of the pain	12.9	5	—	—	12.9	$\chi^2 = 100.33, p < .001$, Cramer's V = .26
5. It is occasional and does not last	38.4	1	2.5	5	35.9	$\chi^2 = 287.66, p < .001$, Cramer's V = .44
6. Medication can take my pain away completely	18.1	4	—	—	18.1	$\chi^2 = 145.11, p < .001$, Cramer's V = .32
7. My pain is because of my current or past work	10.1	8	1.2	7	8.9	$\chi^2 = 53.03, p < .001$, Cramer's V = .19
8. My pain is because of exercise	10.9	7	5.8	2	5.1	$\chi^2 = 16.68, p < .001$, Cramer's V = .10
9. My pain was caused by a recent injury or infection	5.7	10	7.0	1	1.3	$\chi^2 = 1.1, p = .301$, Cramer's V = .03
10. My pain is minor and infrequent	26.4	3	3.3	4	23.1	$\chi^2 = 152.58, p < .001$, Cramer's V = .33
New categories:						
Pain location			49.5			
Back	—		19.8			
Head	—		12.8			
Leg, knee, and foot	—		8.2			
Lower body, hips, and pelvis	—		5.9			
Arms and upper body	—		1.8			
Other (below 1%)	—		1.1			
Pain symptoms	—		18.9			
Stabbing	—		8.8			
Dragging	—		2.6			
Palpatory	—		2.5			

(continued)

Table 3. (continued)

Categories	Closed (N = 706)		Open (N = 729)	
	%	Rank	%	Rank Difference
Radiating pain	—		1.9	
Burning	—		0.8	
Stiffness	—		0.8	
Other pain symptoms (dull, pulsing, and pounding)	—		1.6	
Consequences of pain	—		2.6	
Limited walking ability	—		1.9	
Limited ability to work	—		0.4	
No limitations	—		0.3	
Causes of pain; not code 7, 8, 9	—		18.0	
Illness, disease, condition specified	—		11.8	
Other (e.g., weather, age, birth, menopause)	—		3.2	
Surgery	—		2.2	
Strain, overload	—		0.8	
Pain intensity (bearable, changing), not code 2, 3, 4, 6, 10	—		1.9	
Pain duration (frequent, acute/sudden pain), not code 1, 5, 10	—		1.6	

Note: For all χ^2 -tests, degrees of freedom = 1; N = 1,435. Subcategories are indented and values of subcategories are in italics. The sum of the subcategories is presented in bold type.

or “10. My pain is minor and infrequent.” *Causes of pain* was covered by the three provided response options “7. My pain is because of my current or past work,” “8. My pain is because of exercise” and “9. My pain was caused by a recent injury or infection,” but it was extended by subcategories such as “illness, disease” (in which the condition was specified by respondents), “surgery,” or “strain/overload.”

In Experiment 3, the closed probe following the two questions on self-rated physical activities provided 10 physical activities. Coding of the open-ended probe responses led to the development of four new categories referring to different activities and one category including responses such as “I am not able to do sports” or “I do only light activities.” Slightly less than 10 percent of respondents answering the closed format chose to type their own answer in the text field provided in addition to the category “other” (9.4 percent and 8.3 percent, respectively). Those responses were coded both into the new categories as well as into more diverse subcategories of the provided sets of categories. The main category “cycling, swimming, and other aerobic exercises,” for instance, was supplemented with subcategories such as walking, aerobic, and inline skating. The main category “walking outside of work” could further be differentiated into “dog walking” and “climbing stairs.” To avoid overloading Tables 4 and 5, these fine-grained categories are not shown separately.

Patterns of Interpretation (Response Distributions and Ranking of Themes)

We expected that response distributions would be different across open-ended and closed probes. Hypothesis 2 can be confirmed for all experiments. Hypothesis 3, stating that the most often mentioned categories will be identical across probe formats, can partially be confirmed for the probe following the self-rated health question (Experiment 1) and for both probes following the questions on light/moderate and vigorous activities (Experiment 3), but not for the pain probe (Experiment 2). When comparing the categories that appeared in both formats, Tables 2–5 show substantial differences in the frequency distributions. Predefined response options in the closed format are selected much more frequently than they were mentioned in the open format. Taking a closer look at Experiment 1, the self-rated health probe shows that out of the nine response options provided only two (“ADL without assistance” and “sleep”) are selected by less than 20 percent of respondents in the closed format. In contrast, only one category is selected by more than 20 percent in the open-ended format (“health problems or conditions”). Both

Table 4. Percentage and Ranking of Codes Selected/Mentioned (Patterns of Interpretation) for the Light/Moderate Physical Activity Question.

Categories	Closed (N = 883)		Open (N = 877)		Rank Difference	
	%	Rank	%	Rank		
1. Running or jogging	34.2	4	0.1	16.2	3	$\chi^2 = 74.65, p < .001$, Cramer's $V = .21$
2. Hiking	19.9	7	—	2.3	10	$\chi^2 = 136.60, p < .001$, Cramer's $V = .28$
3. Walking as part of your job	26.2	5	—	0.5	11	$\chi^2 = 250.77, p < .001$, Cramer's $V = .38$
4. Walking outside of work	35.0	2	1.5	19.2	2	$\chi^2 = 54.76, p < .001$, Cramer's $V = .18$
5. Yardwork or cleaning your home	34.2	3	0.7	11.2	5	$\chi^2 = 130.26, p < .001$, Cramer's $V = .28$
6. Working out with exercise equipment	21.0	6	0.5	6.6	7	$\chi^2 = 75.48, p < .001$, Cramer's $V = .21$
7. Lifting weights	13.0	8	—	5.3	9	$\chi^2 = 31.41, p < .001$, Cramer's $V = .14$
8. Cycling, swimming, or other aerobic exercises	35.8	1	1.2	24.4	1	$\chi^2 = 26.72, p < .001$, Cramer's $V = .12$
9. Yoga or stretching	11.3	10	1.9	10.7	6	$\chi^2 = .15, p = .702$, Cramer's $V = .01$
10. Playing a sport	5.4	11	1.5	5.6	8	$\chi^2 = .04, p = .839$, Cramer's $V = .00$
11. Other	11.3	9	9.4 ^a	(12.1) ^b	4	
New categories:						
Sports/exercise, not specified			0.5	8.9		
Sports medicine			0.3	1.3		
Childcare			0.6	1.0		
Other activities (artistic/hobbies/sex)			0.4	0.8		
Not possible/only light activities			—	0.1		

^aSum of all the themes mentioned in the open-ended textbox next to "other"; ^bSum of all "new categories" in the open-ended format, equivalent to "other" in the closed format.

Table 5. Percentage and Ranking of Codes Selected/Mentioned (Patterns of Interpretation) for the Vigorous Physical Activity Question.

Categories	Closed (N = 883)			Open (N = 877)		
	%	Rank	Open „Other”	%	Rank	Diff.
1. Running or jogging	30.8	1	—	17.0	1*	13.8 $\chi^2 = 46.14, p < .001$, Cramer's V = .16
2. Hiking	13.0	8	—	2.5	10	10.5 $\chi^2 = 67.77, p < .001$, Cramer's V = .20
3. Walking as part of your job	12.8	9	—	0.2	11	12.6 $\chi^2 = 113.82, p < .001$, Cramer's V = .25
4. Walking outside of work	14.5	6	0.2	2.9	9	11.7 $\chi^2 = 75.18, p < .001$, Cramer's V = .21
5. Yardwork or cleaning your home	18.5	4	1.6	7.3	6	11.2 $\chi^2 = 48.80, p < .001$, Cramer's V = .17
6. Working out with exercise equipment	24.9	3	0.6	9.1	4	15.8 $\chi^2 = 77.61, p < .001$, Cramer's V = .21
7. Lifting weights	15.9	5	—	7.5	5	8.3 $\chi^2 = 29.54, p < .001$, Cramer's V = .13
8. Cycling, swimming, or other aerobic exercises	25.0	2	1.0	17.0	1*	8.0 $\chi^2 = 17.12, p < .001$, Cramer's V = .10
9. Yoga or stretching	6.7	10	1.4	4.8	8	1.9 $\chi^2 = 2.91, p = .088$, Cramer's V = .04
10. Playing a sport	5.7	11	1.9	6.8	7	-1.2 $\chi^2 = 1.04, p = .307$, Cramer's V = .02
11. Other	13.5	7	8.3^a	(11.9) ^b	3	
New categories:						
Sports/Exercise, not specified			0.1	9.6		
Sports medicine			0.1	0.3		
Childcare			0.1	—		
Other activities (artistic/hobbies/sex)			0.5	0.2		
I don't do sports/not possible/only light activities			0.4	1.7		

Note: * = shared Rank #1.

^aSum of all the themes mentioned in the open-ended textbox next to “other”; ^bSum of all “new categories” in the open-ended format; equivalent to “other” in the closed format.

formats have in common that the option “health problems or conditions” was the most frequently mentioned theme (see Table 2).

For Experiment 2, the probe following the self-rated pain question, the theme “My pain was caused by a recent injury or infection” is the only one which is mentioned by a comparable number of respondents across formats. All other categories are selected significantly more often in the closed format than they are mentioned spontaneously in the open-ended probe (see Table 3 for chi-square tests). Two categories (“When I get my mind on other things, I’m not aware of the pain” and “Medication can take my pain away completely”) are not mentioned at all in the open-ended format. Instead, almost half of the respondents name the location of their pain in the open-ended format. Almost 20 percent of respondents name pain symptoms (18.9 percent) and causes of pain (18.0) other than those already provided as additional categories to describe their pain. The ranking of categories for the pain probe shows that all ranks differ although three response options only differ by one rank each (“2. Sometimes I’m in a lot of pain and sometimes it’s not so bad”; “10. My pain is minor and infrequent”; “7. My pain is because of my current or past work”).

Considering the two probes on light/moderate and vigorous activities in Experiment 3, the differences across formats are not quite as large as in the first two experiments. For both light and vigorous activities, the ranking of the first two categories is the same: The categories “cycling, swimming or other aerobic categories” and “walking outside of work” were selected/mentioned by the largest amount of respondents both in the closed and the open-ended format in response to the probe asking for light/moderate activities. When considering the probe looking into vigorous activities, the categories “running or jogging” as well as “cycling, swimming, and other aerobic exercises” received the same rank, respectively. With regard to the frequency with which an activity is mentioned, the results reveal differences across formats. “Running or jogging” is selected by about one third of all respondents when presented as closed response option, while 16 percent and 17 percent, respectively, mention it in the open-ended format. The largest differences between respondents mentioning a theme regarding light/moderate activities occurs for the category “walking as part of your job,” which is selected by 26 percent in the closed format and by only 0.5 in the open-ended format. It needs to be noted that this category was included as a kind of a control category in the closed probe as the actual survey question only asked for leisure time activities. It is therefore not surprising that this category was only selected in the closed format. Overall, the differences range between 8 and 26 percentage points. However, for the two categories: “yoga or

stretching” and “playing a sport,” there were rather small differences with less than 1 percentage point that were statistically not significant. Comparing the frequency distribution for the probe asking for vigorous activities, the differences range from 1.2 percent points for the response option “playing a sport” up to 16 percent points for the response option “working out with exercise equipment.” For all except one, the categories offered in the closed format are used more frequently than in the open-ended format. Similar to the patterns found for light activities, differences were not significant for the categories “yoga or stretching” and “playing a sport.”

Number of Themes per Respondent

As stated in Hypothesis 4, we expected that respondents in the closed format would on average select more categories than respondents answering the open-ended format will mention themes. Hypothesis 4 can be confirmed, except for the probe following the question on pain in Experiment 2. The number of themes provided by respondents differed significantly for the probe following the self-rated health question with 2.5 in the closed and 1.3 in the open-ended format, one-tailed $t(777) = 13.59, p < .001, d = .98$, and for both probes on light/moderate and vigorous self-rated physical activities, light: 2.4 vs. 1.5 themes; $t(1560) = 14.97, p < .001, d = .76$; vigorous: 1.8 vs. 1.3 themes; $t(1517) = 9.13, p < .001, d = .50$. For the probe to describe self-rated pain, no significant differences in number of themes were found with 1.8 versus 1.7 themes, $t(1345) = 1.56, p < .059, d = .28$.

Nonresponse

We expected that open-ended probes would be more affected by probe nonresponse than closed probes. Since the closed probes were check-all-that-apply formats, nonresponse occurred if no option was selected. For the open-ended question, we differentiate between substantive, codable answers, and nonresponse. Confirming Hypothesis 5, the analyses show that for all four comparisons, the amount of nonsubstantive answers is significantly higher for the open-ended probes than for the closed probes, and these differences have medium effect sizes. The amount of nonresponse answers ranges from 0.1 percent to 1.8 percent in the close-ended format, while it is between 9.1 percent and 11.5 percent in the open-ended format [self-rated health: 0.2 percent vs. 11.1 percent, $\chi^2(1, N = 856) = 45.71, p < .001$, Cramer's $V = .23$; self-rated pain: 0.1 percent vs. 11.1 percent, $\chi^2(1, N = 1,435) = 80.10, p < .001$, Cramer's $V = .24$; light/

moderate activities: 0.5 percent vs. 9.2 percent, $\chi^2(1, N = 1,760) = 73.85, p < .001$, Cramer's $V = .20$; vigorous activities: 1.8 percent vs. 11.5 percent, $\chi^2(1, N = 1,760) = 66.77, p < .001$, Cramer's $V = .19$].

Impact of Sociodemographic Characteristics

We expected the closed probes to be easier for lower educated respondents, resulting in less nonresponse and more themes than the open-ended probes. Hypothesis 6 cannot be confirmed. To evaluate effects of sociodemographic characteristics and probe format on mean number of themes mentioned, we ran two-way ANOVAs for sex and education.

The analysis revealed a significant main effect for education for the probes following self-rated health, $F(5, 776) = 8.0, p = .004, \eta^2 = .02$, and light/moderate activities, $F(5, 1512) = 7.1, p < .001, \eta^2 < .01$, but no significant interaction between education and probe format, $F(5, 776) = 1.0, p = .366$; $F(5, 1512) = .03, p = .973$. For the probes following self-rated pain and vigorous activities, no significant main effect for education was found, $F(3, 1434) = 2.2, p = .115$; $F(3, 1501) = 0.5, p = .598$. In the probe following on self-rated pain, a significant main effect of sex was found, $F(3, 1434) = 8.3, p < .004, \eta^2 < .01$, with women mentioning more themes than men. No significant interaction effects of sex with probe format were found in all four probes. We also ran three-way ANOVAs with sex, education, and probe format as independent variables. No significant interaction effects were found.

As probe nonresponse in the closed format was very low in most of the experiments (with a maximum of $n = 16$), valid comparisons across formats were not suitable. A closer examination of sociodemographic characteristics of respondents providing nonresponse in the open-ended probe format only revealed that men provided significantly more nonresponse than women in all four probes asked. The amount of nonresponse differed significantly between the lowest and the highest educational group in three out of four probes (except the probe following on light activities; the results of the chi-square tests can be found in Table A1 in the Appendix, which can be found at <http://smr.sagepub.com/supplemental/>). With regard to age, the findings are mixed: There were no significant differences for the probes on self-rated health and pain, but there were significant differences for the probes following the questions on light/moderate and vigorous activities. In the latter two probes, the amount of nonresponse in the oldest age group (aged 50+) was significantly lower than in the two younger age groups (see Table A1).

Discussion

The novel method of embedding closed probes into web surveys, as proposed by Scanlon (2020a), seeks to provide information about the “patterns of interpretation a respondent uses when answering a survey item” (Scanlon 2020a:446). To be able to replace open-ended probes with closed probes, it is important that both forms yield identical or at least similar results and that they reflect the same cognitive processes that respondents are going through while answering the survey questions. In this study, we compared open-ended and closed probes in the context of web probing with regard to substantive content and nonresponse, the latter being an important indicator of data quality. The comparisons we made revealed statistically significant and substantially important differences in distributions between open-ended and closed probe formats. Our results show that both formats do not provide comparable results with regard to nonresponse, the coverage of themes, the patterns of interpretation, and the number of themes per respondent. Comparing nonresponse, our results show higher rates of nonresponse in the open-ended format, which indicates a higher response burden in the open-ended format compared to the closed format. In contrast, the closed probes are answered by almost all respondents.

With regard to substantive issues, the open-ended probe responses could be coded in a more detailed way, resulting in new categories and more specific subcategories than the provided closed options. This indicates that the closed format cannot cover the full picture of what respondents associated with the survey question and is thus no substitute for open-ended probes, at least if researchers are interested in learning about the breadth of possible interpretations of an item. In terms of response distributions, and hence patterns of interpretation, it remains unclear which results are more valid and reflect the response process more accurately. Two explanations seem reasonable: First, respondents in the closed format are more likely to mention themes that they did not think of while answering the survey question but are reminded of when they are displayed as explicit response options—this will result in more themes per respondent. Second, respondents in the open-ended format may not be willing to write down all of their thoughts due to the perceived burden—this will result in fewer themes per respondent. Answering probes in a closed or open-ended format could also generally be perceived as a different response task: Due to the “select-all-that-apply” instruction in the closed format, respondents could have understood the probe as “what applies to you” instead of “what were you thinking of when answering the previous question.” Hence, they may have selected statements

or categories that apply somehow to their life situation but that they did not initially think of. For example, in asking “what did you think of” in assessing health in general and presenting a range of possibilities (e.g., diet, nutrition), it is possible that respondents may choose categories that they deem appropriate but that they were not, in fact, thinking of when they answered the initial question. Thus, the closed approach might be useful to learn about how respondents think about a concept of interest in general (e.g., “Which of the these do you think are important for your health?” followed by a list of predetermined response options). However, this approach may represent a fundamentally different task than the one posed by using an open-ended format, for which it is more likely that respondents will respond with what they were thinking of, as instructed.

Based on the advantages and disadvantages of each probe format and our findings, it becomes clear that both the wording of the probe and thereby the range of interpretations as well as the content of response options play a decisive role. Both are subject to the decisions made by the researchers when developing the probes. If an open-ended probe is worded rather broadly and the construct under investigation contains many subdimensions (as for example in the case of pain), this combination could at least partly explain the discrepancies between the frequency distributions. Respondents in the closed format following the pain probe were guided by the response suggestions, which contained a specified range of aspects (e.g., experience, intensity, cause), while respondents in the open-ended format had no guidance at all which direction the probe was targeting, resulting in a large number of responses indicating the pain location. As a practical recommendation, it can be concluded that probes must be formulated as precisely as possible, in particular, in the open-ended format and in the context of web probing since there is neither an interviewer who could guide the response process or follow-up on issues raised by respondents nor any predefined response suggestions. In this context, it should be considered that comprehension probes asking for the understanding of a specific term (“What does the term X mean to you?”) are more effective than probes asking for elaborative information, such as “Can you please explain your answer a little further?” (cf. Foddy 1998).

With regard to the wording of the closed response categories, the following observations could be made: Some categories were indistinguishable (e.g., “It is occasional and does not last”; “My pain is minor and infrequent”); and some categories were also biased (e.g., asking about infrequent pain, but not about frequent pain). It could further be observed that the more specific the closed categories were (e.g., pain because of *recent* infection), the smaller

the number of respondents who were able to choose this category. Hence, particular attention should also be paid to the wording and the distinctiveness of the closed probe options. In particular, the process of “translating” verbal data collected in cognitive interviews into meaningful closed response options requires aggregating the information into broader categories and covering opposites as well (e.g., occurrence of both infrequent and frequent pain).

Also, when providing sets of response categories, it is possible that one additional cognitive interview could yield a significant new issue, which may then result in an additional response category. At the same time, it is not possible to expand the list of response options infinitely. Researchers have to decide where to cut off the list of closed categories, which will automatically be accompanied by a certain loss of information. With a longer list of provided response options, response burden increases, which might result in increased nonresponse rates also for closed probes.

How many response categories should be offered will also depend on the specific research objective. There are research interests for which the number and type of provided response categories can be determined in advance. Whether respondents have positive or negative associations with a term is one such research question for which Scanlon (2020a) gives an example. Respondents were asked to state whether they “would consider everything being an effort to be a good thing, or a bad thing” with three response options: good thing, bad thing, neither good nor bad (p. 440).

To conclude, open-ended and closed formats both have strengths and weaknesses and it is less a decision of one approach over the other. Instead, researchers have to select the appropriate approach for their research question. It seems that closed-ended probes are likely more useful when the researchers have a particular (and relatively narrow) hypothesis or prediction that they would like to investigate and can formulate a probe and the response categories to examine that issue specifically (e.g., When thinking about health in general, do you think about your physical health, mental health, or both?) However, when the objective is more general or when researchers are interested in the full breadth of interpretations, implementing open-ended probes seems to be more appropriate.

The study has some limitations that call for further research. First, the four survey questions under investigation were solely behavioral questions. Future research could investigate whether the two probe formats are more comparable when attitudinal or opinion questions are probed. Also, the questionnaire contained mostly aspects of health, which could be of varying relevance and interest to subgroups of respondents (e.g., elderly people). Second, we ourselves were not involved in developing the closed probes;

hence, we cannot share information and experience on what it means to transform cognitive interviewing responses into closed web probes. We did test the closed probes in cognitive interviews in German. There were no further additions to the categories following the German cognitive interviews. Ultimately, we adhered to the closed probes provided by our American colleagues to ensure comparability with the U.S. data (the intercultural component was of prime importance in this study).

In how far it is sufficient to adopt probes and response options provided by a collaborator living in another cultural context or whether they should generally be developed in the respective cultural context should be further investigated. The latter would have implications for the usability of closed probes in an intercultural context. As cross-cultural studies aim for cross-cultural comparability of measurements and similar understanding of concepts, the method of embedding closed probes could be helpful to determine whether the patterns of interpretation are similar or different. This would make it possible to identify prior to fielding the survey whether a certain question is associated with substantial differences in understanding its concept across the countries.

Finally, more research is needed regarding the effect of the number of response options provided to present a complete picture and regarding an alleged increase in response burden when implementing closed probes.

Acknowledgment

The authors are grateful to Paul Scanlon, Kristen Miller (National Center of Health Statistics) and Bridget Reynolds for sharing their expertise as well as the probing questions used in the National Center for Health Statistics' Research and Development Survey (RANDS) and, additionally, to Paul Scanlon for valuable comments on an earlier version of the manuscript.


Declaration of Conflicting Interests


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

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID iDs

Cornelia E. Neuert  <https://orcid.org/0000-0001-9855-5618>

Katharina Meitinger  <https://orcid.org/0000-0001-8160-556X>

Supplemental Material

The supplemental material for this article is available online.

Notes

1. For example: For the question on general health (“Would you say your health in general is excellent, very good, good, fair, or poor?”), “in-person” cognitive interviews revealed different patterns of interpretation, which were then used in the following embedded probe: “When you answered the previous question about your health, what did you think of?” with the following seven response options: “My diet and nutrition. My exercise habits. My drinking habits. My health problems or conditions. The amount of times I seek health care. The amount of pain or fatigue that I have. My conversations with my doctor” (Scanlon 2020a:432).
2. Recent studies have come to experiment with voice answers (compared to text answers; Gavras and Höhne 2020; Revilla et al. 2020). This line of research will certainly be extended to cognitive probes in the future, too.
3. As described in Scanlon (2020a), the closed probes were designed based on the findings from three iterative rounds of cognitive interviews. The findings from the “in-person” cognitive interviews were used to uncover the patterns of interpretation respondents had in mind while answering the target survey questions. These patterns were used to develop a response scheme which then served as the basis for the closed probes whereby each pattern typically became one of the response categories. In the Scanlon (2020) study, the aim of developing and embedding the closed probes was to determine the frequency of each of these patterns across the survey population, that is, to understand which of the patterns respondents used when answering the survey item under evaluation (p. 432).

References

- Andrews, Mark. 2005. “Who Is Being Heard? Response Bias in Open-Ended Responses in a Large Government Employee Survey.” Pp. 3760-66 in *60th Annual Conference of the American Association for Public Opinion Research*. Miami Beach, FL: Methods A-ASoSR.
- Beatty, Paul C. and Gordon B. Willis. 2007. “Research Synthesis: The Practice of Cognitive Interviewing.” *Public Opinion Quarterly* 71(2):287-311.
- Behr, Dorothee, Wolfgang Bandilla, Lars Kaczmirek, and Michael Braun. 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(4): 524-33.
- Behr, Dorothee, Katharina Meitinger, Michael Braun, and Lars Kaczmirek. 2017. “Web Probing—Implementing Probing Techniques from Cognitive Interviewing

- in Web Surveys with the Goal to Assess the Validity of Survey Questions.” in *GESIS – Survey Guidelines*, Mannheim: GESIS – Leibniz-Institute for the Social Sciences. Accessed 6th July 2021 (https://doi.org/10.15465/gesis-sg_en_023).
- Borg, Ingwer and Cornelia Züll. 2012. “Write-in Comments in Employee Surveys.” *International Journal of Manpower* 33(2):206-20.
- Bradburn, Norman. 1978. “Respondent Burden.” in *Proceedings of the Survey Research Methods Section of the American Statistical Association*, 35:40. Alexandria, VA: American Statistical Association.
- Callegaro, Mario and Charles DiSogra. 2008. “Computing Response Metrics for Online Panels.” *Public Opinion Quarterly* 72(5):1008-32.
- Converse, J. M. and Stanley Presser. 1986. *Survey Questions: Handcrafting the Standardized Questionnaire*. Thousand Oaks, CA: Sage.
- Crawford, Scott D., Mick P. Couper, and Mark J. Lamias. 2001. “Web Surveys: Perceptions of Burden.” *Social Science Computer Review* 19(2):146-62.
- Denscombe, Martyn. 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(3):359-68.
- Dillman, Don A., Jolene D. Smyth, and Leah Melani Christian. 2009. *Internet, Mail, and Mixed Method Surveys: The Tailored Design Method*. Hoboken, NJ: John Wiley.
- Edgar, Jennifer, Joe Murphy, and Michael Keating. 2016. “Comparing Traditional and Crowdsourcing Methods for Pretesting Survey Questions.” *SAGE Open* 6(4). Accessed 6th July 2021 (<https://doi.org/10.1177/2158244016671770>).
- Foddy, William. 1998. “An Empirical Evaluation of In-Depth Probes Used to Pretest Survey Questions.” *Sociological Methods & Research*, 27(1):103-33.
- Friborg, Oddgeir and Jan H. Rosenvinge. 2013. “A Comparison of Open-Ended and Closed Questions in the Prediction of Mental Health.” *Quality & Quantity* 47(3): 1397-411.
- Galesic, Mirta. 2006. “Dropouts on the Web: Effects of Interest and Burden Experienced during an Online Survey.” *Journal of Official Statistics* 22(2):313-28.
- Garbarski, Dana, Jennifer Dykema, Kenneth D. Croes, and Dorothy F. Edwards. 2017. “How Participants Report Their Health Status: Cognitive Interviews of Self-Rated Health across Race/Ethnicity, Gender, Age, and Educational Attainment.” *BMC Public Health* 17(1):771. Accessed 6th July 2021 (<https://doi.org/10.1186/s12889-017-4761-2>).
- Gavras, Konstantin and Jan Karem Höhne. 2020. “Evaluating Political Parties: Criterion Validity of Open Questions with Requests for Text and Voice Answers.” *International Journal of Social Research Methodology*. Accessed 6th July 2021 (<https://doi.org/10.1080/13645579.2020.1860279>).
- Geer, John G. 1991. “Do Open-Ended Questions Measure “Salient” Issues?” *Public Opinion Quarterly* 55(3):360-70.

- Groves, R. M., F. J. Fowler Jr, M. P. Couper, J. M. Lepkowski, E. Singer, and , & R. Tourangeau (2011). *Survey methodology*. John Wiley & Sons.
- Guest, Greg, Arwen Bunce, and Laura Johnson. 2006. "How Many Interviews are Enough? An Experiment with Data Saturation and Variability." *Field Methods*, 18(1):59-82.
- Harkness, Janet A. 2003. "Questionnaire Translation." Pp. 35-56 in *Cross-Cultural Survey Methods*, edited by J. Harkness, F. J. R. van de Vijver, and P. Mohler. Hoboken, NJ: Wiley.
- Holland, Jennifer L. and Leah Melani Christian. 2009. "The Influence of Topic Interest and Interactive Probing on Responses to Open-Ended Questions in Web Surveys." *Social Science Computer Review* 27(2):196-212.
- Knapp, Frank and Martin Heidingsfelder. 2001. "Drop-Out Analysis: Effects of the Survey Design." Pp. 221-30 in *Dimensions of Internet Science*, edited by U. D. Reips and M. Bosnjak. Lengerich: Pabst Science Publishers.
- Lee, Sunhee, Colleen McClain, Dorothee Behr, and Katharina Meitinger. 2020. "Exploring Mental Models Behind Self-Rated Health and Subjective Life Expectancy through Web Probing." *Field Methods* 32(3):309-26.
- Lenzner, Timo and Cornelia E. Neuert. 2017. "Pretesting Survey Questions via Web Probing—Does It Produce Similar Results to Face-to-Face Cognitive Interviewing?" *Survey Practice* 10(4). Accessed 6th July 2021 (<https://doi.org/10.29115/SP-2017-0020>).
- Lenzner, Timo and Cornelia E. Neuert. 2019. "What Makes a 'Good' Web Probing Respondent." Paper presented at the *8th Conference of the European Survey Research Association*, July, Zagreb (Croatia).
- Meitinger, Katharina. 2017. "Necessary but Insufficient: Why Measurement Invariance Tests Need Online Probing as a Complementary Tool." *Public Opinion Quarterly* 81(2):447-72.
- Meitinger, Katharina and Dorothee Behr. (2016). "Comparing Cognitive Interviewing and Online Probing: Do They Find Similar Results?" *Field Methods* 28(4): 363-80.
- Meitinger, Katharina, Michael Braun, and Dorothee Behr. 2018. "Sequence Matters in Online Probing: The Impact of the Order of Probes on Response Quality, Motivation of Respondents, and Answer Content." *Survey Research Methods* 12(2):103-20.
- Meitinger, Katharina and Lars Kaczmirek. 2018. "Identifying of and Dealing with Item Nonresponse in Open-Ended Questions in a Cross-National Context." [pdf slides]. Retrieved August 26, 2020 (<https://csdiworkshop.org/past-events/2018-csdi-workshop/2018-presentations/>)
- Oudejans, Marije and Leah Melani Christian. 2010. "Using Interactive Features to Motivate and Probe Responses to Open-Ended Questions." Pp. 215-44 in *Social*

- and Behavioral Research and the Internet, edited by M. Das, P. Ester, and L. Kaczmarek. New York: Routledge.
- Reja, Urša, Katja Lozar Manfreda, Valentina Hlebec, and Vasja Vehovar. 2003. "Open-Ended vs. Close-Ended Questions in Web Questionnaires." *Developments in Applied Statistics* 19(1):159-77.
- Revilla, Melanie, Mick P. Couper, Oriol J. Bosch, and Mario Asensio. 2020. "Testing the Use of Voice Input in a Smartphone Web Survey." *Social Science Computer Review* 38(2):207-24.
- Scanlon, Paul J. 2019. "The Effects of Embedding Closed-Ended Cognitive Probes in a Web Survey on Survey Response." *Field Methods* 31(4): 328-43.
- Scanlon, Paul J. 2020a. "Using Targeted Embedded Probes to Quantify Cognitive Interviewing Findings." Pp. 427-49 in *Advances in Questionnaire Design, Development, Evaluation and Testing*, edited by P. Beatty, D. Collins, L. Kaye, J. Padilla, G. Willis, and A. Wilmot. Hoboken, NJ: Wiley.
- Scanlon, P. (2020b). Cognitive evaluation of the National Center for Health Statistics' 2018 Research and Development Surveys. National Center for Health Statistics Q-Bank. 2020. <https://www.cdc.gov/qbank/Reports.aspx#/Reports/1203>. Accessed date [07/06/2021]
- Schmidt, Katharina, Tobias Gummer, and Joss Roßmann. 2020. "Effects of Respondent and Survey Characteristics on the Response Quality of an Open-Ended Attitude Question in Web Surveys." *Methods, Data, Analyse* 14(1):3-34.
- Scholz, Evi and Cornelia Züll. 2012. "Item Non-Response in Open-Ended Questions: Who Does Not Answer on the Meaning of Left and Right?" *Social Science Research* 41(6):1415-28.
- Schuman, Howard. 1966. "The Random Probe: A Technique for Evaluating the Validity of Closed Questions." *American Sociological Review* 31(2):218-22.
- Schuman, Howard and Stanley Presser. 1979. "The Open and Closed Question." *American Sociological Review* 44:692-712.
- Schuman, Howard and Stanley Presser. 1981. *Questions and Answers in Attitudes Surveys*. San Diego, CA: Academic Press.
- Schwarz, Norbert. 1999. "Self-Reports: How the Questions Shape the Answers." *American Psychologist*, 54(2):93-105.
- Silber, Henning, Cornelia Züll, and Steffen M. Kuehnel. (2020). "What Can We Learn from Open Questions in Surveys? A Case Study on Non-Voting Reported in the 2013 German Longitudinal Election Study." *Methodology* 16(1):41-58.
- Smyth, Jolene D., Don A. Dillman, Leah Melani Christian, and Mallory Mcbride. 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(2):325-37.

- Stern, Michael J., Don A. Dillman, and Jolene D. Smyth. 2007. "Visual Design, Order Effects, and Respondent Characteristics in a Self-Administered Survey." *Survey Research Methods* 1(3):121-38.
- Tourangeau, Roger. 1984. "Cognitive Science and Survey Methods." Pp. 73-100 in *Cognitive Aspects of Survey Methodology: Building a Bridge Between Disciplines*, edited by T. B. Jabine, M. L. Straf, J. M. Tanur, and R. Tourangeau. Washington, DC: National Academy Press.
- Willis, Gordon B. (2005). *Cognitive Interviewing: A Tool for Improving Questionnaire Design*. New York: Sage.
- Züll, Cornelia. 2016. "Open-Ended Questions." *GESIS Survey Guidelines*. Mannheim, Germany: GESIS – Leibniz Institute for the Social Sciences. doi: 10.15465/gesis-sg_en_002.

Author Biographies

Cornelia E. Neuert is a postdoctoral researcher and head of the team Questionnaire Design & Evaluation at GESIS—Leibniz Institute for the Social Sciences. Her research interests include questionnaire design, questionnaire evaluation, and data quality.

Katharina Meitinger is an assistant professor at Utrecht University. Her research focuses on cross-cultural survey methodology and measurement quality, in particular mixed-method designs and the method of web probing.

Dorothee Behr is a senior researcher and head of the team Cross-cultural Survey Methods at GESIS—Leibniz Institute for the Social Sciences. Her research interests focus on web probing, questionnaire translation, and comparability of cross-cultural survey data.