*Journal of Survey Statistics and Methodology (2024) **00**, 1–23*

INNOVATING WEB PROBING: COMPARING WRITTEN AND ORAL ANSWERS TO OPEN-ENDED PROBING QUESTIONS IN A SMARTPHONE SURVEY

TIMO LENZNER*
JAN KAREM HÖHNE
KONSTANTIN GAVRAS

Cognitive interviewing in the form of probing is key for developing methodologically sound survey questions. For a long time, probing was tied to the laboratory setting, making it difficult to achieve large sample sizes and creating a time-intensive undertaking for both researchers and participants. Web surveys paved the way for administering probing questions over the Internet in a time- and cost-efficient manner. In so-called web probing studies, respondents first answer a question and then they receive one or more open-ended questions about their response process, with requests for written answers. However, participants frequently provide very short or no answers at all to open-ended questions, in part because answering questions in writing is tedious. This is especially the case when the web survey is completed via a smartphone with a virtual on-screen keypad that shrinks the viewing space. In this study, we examine whether the problem of short and uninterpretable answers in web probing studies can be mitigated by asking respondents to complete the web survey on a smartphone and to record their answers via the built-in microphone. We conducted an experiment in a smartphone survey ($N = 1{,}001$), randomizing respondents to different communication modes (written or oral) for answering two

Timo Lenzner is Senior Researcher at GESIS—Leibniz Institute for the Social Sciences, Survey Design and Methodology, P.O. Box 12 21 55, 68072 Mannheim, Germany. Jan Karem Höhne is Junior Professor for Computational Survey and Social Science at Leibniz University Hannover, Hannover, Germany; German Centre for Higher Education Research and Science Studies (DZHW), Lange Laube 12, 30159 Hannover, Germany; RECSM-Universitat Pompeu Fabra, Barcelona, Spain. Konstantin Gavras is Data Scientist at Nesto Software GmbH, Tullastraße 58, 76131 Karlsruhe, Germany.

The authors would like to thank Marisa Finzer and Joanna Scholl (GESIS) for their support in transcribing and coding respondents' answers to the open-ended probes. They also thank Joshua Claaßen (DZHW) for his support during the writing of this article. This study design and analysis was not preregistered.

*Address correspondence to Timo Lenzner, GESIS—Leibniz Institute for the Social Sciences, Survey Design and Methodology, P.O. Box 12 21 55, 68072 Mannheim, Germany; E-mail: timo.lenzner@gesis.org

https://doi.org/10.1093/jssam/smae031

comprehension probes about two questions on national identity and citizenship. The results indicate that probes with requests for oral answers produce four to five times more nonresponse than their written counterparts. However, oral answers contain about three times as many words, include about 0.3 more themes (first probing question only), and the proportion of clearly interpretable answers is about 6 percentage points higher (for the first probing question only). Nonetheless, both communication modes result in similar themes mentioned by respondents.

KEY WORDS: Cognitive pretesting; Experiment; Smartphone survey; Survey question design; Voice recording; Web probing.

Statement of Significance

Web surveys are a prevailing data collection method in social science research and many adjacent disciplines. This is accompanied by an increasing body of web- instead of lab probing studies to improve survey question design and data quality on a large-scale level. The growing number of web survey completions through smartphones additionally allows researchers to make web probing studies more human-like by asking probing questions, with requests of oral instead of written answers. Although oral answers collected through the built-in microphone of smartphones provide a promising avenue for survey question design, there are no studies experimentally comparing web probing studies that employ written and oral answers to provide empirical-driven state-of-the-art recommendations. This study attempts to close this research gap by reporting the results of an experiment incorporated in a large-scale smartphone survey. Its results contribute to our contemporary knowledge on web probing studies and are of key interest for survey researchers and practitioners.

1. INTRODUCTION AND BACKGROUND

Cognitive pretesting is an essential step in developing high-quality survey questions, that is, questions measuring what they are supposed to measure. At its core, it involves asking respondents questions about how they understood specific terms in survey questions and how they arrived at their answers (Willis 2005; Collins 2015). Such questions about survey questions are commonly referred to as *probes*. Examples of commonly used probes are "What does the term X mean to you?" (comprehension probe), "How did you arrive at that answer?" (process probe), and "How do you know that you went to the doctor three times in the past 12 months?" (recall probe) (Willis and Miller

2011). The aim of asking probes is to reveal respondents' mental processes when answering survey questions (Miller et al. 2014) and determine, for example, whether they interpreted the questions as intended or whether they had difficulties providing an answer (Converse and Presser 1986). To put it differently, probing is a question evaluation method that aims to assess the validity of survey questions by generating self-reports from respondents about their cognitive response processes (Padilla and Benítez 2014).

Traditionally, cognitive pretesting in general and probing in particular have been carried out in lab settings, where specially trained interviewers conduct in-depth cognitive interviews with small samples of participants (Willis 2005). However, with the advent of self-administered web surveys, a new method called *web probing* has emerged, in which the probing techniques from lab-based cognitive interviewing are transferred into a web survey (Behr et al. 2012). In web probing studies, respondents first answer one or more survey questions and then they are presented with probes about these questions. Typically, these probing questions come with an open-ended answer format (i.e., respondents enter a narrative answer in a text field). Web probing has become particularly popular for evaluating questions that are used in self-administered web surveys (Fowler and Willis 2020) and in the context of evaluating questions in cross-national surveys; for example, to explore reasons for the lack of measurement invariance (Meitinger 2017). In addition, web probing is commonly used in the questionnaire development and pretesting phase of surveys (see Hadler et al. 2018; Lenzner et al. 2022).

Compared to lab-based cognitive interviewing, web probing offers several advantages. First and foremost, it allows for faster and less expensive recruitment of respondents and thus the realization of larger samples. This, in turn, allows researchers to better quantify their pretest results (Behr et al. 2012), conduct statistical analyses in addition to qualitative ones (e.g., determine the internal consistency and item-total correlations in item batteries; Schick et al. 2023), and carry out subgroup comparisons (Neuert et al. 2021). Second, web probing facilitates the recruitment of (specific) respondents from different groups, regions, cultures, or countries. Third, the self-administered mode eliminates potential interviewer effects, thereby increasing the quality and comparability of results (Conrad and Blair 2009).

However, the lack of an interviewer asking the questions impedes the possibility to probe for additional information or to follow up on incomplete answers. Probing is limited to the questions previously programmed into the web survey. Moreover, it is difficult to motivate respondents to answer the open-ended probes thoughtfully and in detail while completing the web survey. As a result, web probing studies are often characterized by higher probe nonresponse than lab-based cognitive interviewing studies. In addition, answers are frequently shorter and less clearly interpretable (i.e., overly vague and lacking the necessary details for meaningful analyses; Meitinger and Behr 2016; Lenzner and Neuert 2017; Fowler and Willis 2020). Mobile device use

in web surveys exacerbates these problems, presumably due to device-related issues, such as the virtual on-screen keypad shrinking the viewing space and making it even more tedious to answer a question in an open-ended format (Struminskaya et al. 2015; Lugtig and Toepoel 2016; Revilla and Ochoa 2016). At the same time, the share of respondents completing web surveys on mobile devices has increased substantially over recent years (Revilla et al. 2016; Gummer et al. 2023). This development, paired with a general reluctance of respondents to answer open-ended questions in self-administered web surveys in a comprehensive and detailed manner, poses a threat to the quality of data collected in web probing studies.

A promising way to mitigate the problem of short and uninterpretable answers in web probing studies is to ask respondents to complete the web survey on their smartphones and to record their answers via the built-in microphone. This would transfer the more natural (verbal) communication from lab-based cognitive interviews to the online environment (Schober et al. 2015; Gavras and Höhne 2020; Revilla et al. 2020; Revilla and Couper 2021; Gavras et al. 2022). It resembles the communication features implemented in popular instant messaging services, such as WhatsApp and WeChat (Höhne 2023). By partially simulating everyday conversations, this setting has the potential to encourage respondents to engage in open narrations, thereby allowing researchers to gather rich and nuanced information from them (Gavras and Höhne 2020; Gavras et al. 2022).

In this study, we build on these developments in web survey research and examine whether the problems associated with requests for written answers in web probing studies can indeed be mitigated by asking respondents to provide oral answers in a smartphone survey. Specifically, our research objective is to examine whether oral answers result in higher-quality data in web probing studies than their written counterparts. For this purpose, we conducted a smartphone survey in a German online access panel, asking two questions about the relationship between citizens and state, as well as two comprehension probes about these questions.

2. RESEARCH QUESTIONS AND HYPOTHESES

In this study, we address three research questions (RQs) with corresponding hypotheses derived from theoretical considerations and empirical findings. We outline these RQs and hypotheses in the following paragraphs.

RQ1: Does the communication mode (oral vs. written answers) in web probing studies influence the quality of respondents' answers to the probes?

Previous studies reported higher probe nonresponse in web probing compared to lab-based cognitive interviewing (Meitinger and Behr 2016; Lenzner

and Neuert 2017). In a study by Meitinger and Behr (2016), for example, probe nonresponse ranged from 11 percent to 31 percent for different probes when these were administered in a web probing study, whereas it was almost nonexistent when the same probes were asked in lab-based cognitive interviews. These findings might be due to the self-administered survey mode in web probing: first, it may be difficult for respondents to express their mental processes in a written way, especially if they have low literacy skills (Gavras et al. 2022). Second, there is no interviewer present to motivate respondents to invest the effort required to answer open-ended questions (Lenzner and Neuert 2017). Of course, interviewers are also absent in web probing studies requesting oral answers, so one would expect a similar effect on probe nonresponse in this self-administered setting. Indeed, earlier studies comparing oral and written answers in smartphone surveys have found that many respondents are reluctant to provide oral answers (Revilla et al. 2020; Revilla and Couper 2021; Gavras et al. 2022). For example, Gavras et al. (2022) reported item nonresponse rates of 25 percent to 28 percent for open-ended questions with requests for oral answers, but less than 5 percent for open-ended questions with requests for written answers. Based on these earlier findings, we postulate the following hypothesis:

Hypothesis 1: Probe nonresponse is higher when respondents are asked for oral answers than for written answers.

Earlier studies found that self-administered settings (web probing) produce more uninterpretable answers to open-ended probes than interviewer-administered settings (Meitinger and Behr 2016; Lenzner and Neuert 2017). On the one hand, this could be due to the absence of an interviewer, which makes it hard to clarify the intent of a probe or to follow up on a respondent's answer that is not clearly interpretable. On the other hand, it may be because web probing answers are considerably shorter (and possibly less elaborated) than answers in lab-based cognitive interviews (Meitinger and Behr 2016). Earlier research comparing oral and written answers in smartphone surveys has suggested that oral answers are longer than their written counterparts and are associated with rather open narrations (Gavras et al. 2022). This may be due to several factors, including that answering questions via the built-in microphone of smartphones may be less burdensome than typing them in. In addition, Gavras et al. (2022) argue that requests for written answers are more likely to trigger a memory-based processing (Zaller and Feldman 1992; Tourangeau et al. 2000) and may thus be more intentional and conscious, whereas oral answers are more likely to trigger an online processing (Lodge et al. 1989; McGraw et al. 2003) and may thus be more intuitive and spontaneous. The presumably less burdensome answer delivery associated with oral answers, coupled with an online processing, potentially results in longer answers. Whatever factor or combination of factors is responsible for this finding, it suggests that the longer oral answers may leave less room for

interpretation or less ambiguity about what respondents are trying to express. From these considerations and previous findings, we derive the following two hypotheses:

Hypothesis 2: The amount of not clearly interpretable answers to open-ended probes is lower when respondents are asked for oral answers than for written answers.

Hypothesis 3: Respondents use more words to answer open-ended probes when asked for oral answers than for written answers.

RQ2: Does the communication mode (oral vs. written answers) in web probing studies affect the content or depth of respondents' answers?

An earlier study comparing oral answers in lab-based cognitive interviews with written answers in web probing studies found that the number of themes mentioned by respondents did not differ between the two communication modes (Meitinger and Behr 2016). However, empirical evidence from smartphone surveys indicates that oral answers result in a higher number of themes than their written counterparts (Gavras et al. 2022). One explanation for this finding is that written answers involve an intentional and conscious memory-based processing, but their answer delivery is more tedious and burdensome than the answer delivery associated with oral answers that follow an intuitive and spontaneous online processing. This may prevent respondents from writing down all relevant aspects they think of (Gavras et al. 2022). Therefore, we postulate the following hypothesis:

Hypothesis 4: Respondents mention more themes in open-ended questions when asked for oral answers than for written answers.

Meitinger and Behr (2016) reported some differences in the types of themes being mentioned by respondents in lab-based cognitive interviews (oral answers) and in a web probing setting (written answers). Specifically, some themes were only mentioned by web probing respondents. The authors argue that this might be due to the larger sample sizes in web probing. However, research comparing oral and written answers to open-ended questions in smartphone surveys (Gavras et al. 2022) suggests that it might also be due to the different communication modes in both settings. The latter study reported an overlap of less than 50 percent between themes mentioned in written and oral answers. The authors argue that different cognitive processes (memory-based and online) are at play when respondents provide written (memory-based) and oral answers (online). In correspondence with this reasoning and the earlier findings, we postulate the following hypothesis:

Hypothesis 5: Respondents mention different themes in open-ended questions when asked for oral answers than for written answers.

RQ3: Does the communication mode (oral vs. written answers) influence how respondents evaluate the survey?

Previous research has found that a substantial share of respondents is not willing to participate in smartphone surveys requesting oral answers (Lenzner and Höhne 2022; Höhne 2023). The most often cited reasons for this reluctance were a preference for writing rather than speaking, a preference for completing surveys on a PC rather than a smartphone, and privacy and data security concerns (Lenzner and Höhne 2022). As mentioned above, studies examining the factual willingness to provide oral answers in web surveys consistently attest requests for oral answers to be associated with higher item nonresponse than requests for written answers (Revilla et al. 2020; Gavras et al. 2022). Based on this finding, one might expect that respondents evaluate a survey (request) less positively when asked to provide oral instead of written answers. It is possible that they do not perceive this type of questioning as entertaining and convenient, but as intrusive and complicated. This would not only discourage some respondents from participating in the survey in the first place, but also encourage low-quality answers (i.e., satisficing behavior; Krosnick 1991) among those who start the survey. On the other hand, exempting respondents from the burden of typing in answers to open-ended questions might lead them to perceive a survey as easier and faster to complete, and thus to rate it more positively (e.g., more interesting) and to provide higher-quality responses. It is unclear whether one or the other mechanism has a stronger influence on a survey's evaluation or whether they cancel each other out. Hence, we postulate the following hypothesis:

Hypothesis 6: Respondents do not differ in their evaluation of a web survey's level of interest, difficulty, or length, regardless of whether they are asked for oral or written answers.

3. METHOD

3.1 Data Collection

Data were collected in Germany during November 2021 through the Forsa Omninet Panel. The target population of the Omninet Panel are German-speaking Internet users aged 18–76 who live in Germany. Respondents cannot register themselves (to avoid fake accounts and duplicates) but are invited via a probability-based telephone sample. The survey mode in the Omninet Panel is online. Forsa selected a cross-quota sample from their online panel based on age (born between 1985 and 2003, 1965 and 1984, and 1945 and 1964) and gender (female and male). Additionally, they drew quotas on education (low = graduated from a lower secondary school or less, middle = graduated from an intermediate secondary school, and high = graduated from a college

preparatory school or university). All quotas were selected based on the German Microcensus, which was used as a population benchmark. Data and code are available for replication purposes via Harvard Dataverse (https://doi.org/10.7910/DVN/P9PHUF).

Participants were invited to the web survey via email and were provided with information about the device to be used for participation (smartphone) and a link to the web survey. The first page of the web survey provided an overview of the survey topics and outlined the overall procedure. It also included a statement of confidentiality assuring that the study complies with existing data protection laws and regulations. Prior informed consent for data collection was obtained by Forsa. Respondents received financial compensation worth 1 Euro (in the form of bonus points) from Forsa for their participation.

To limit participation in the web survey to smartphone respondents, we identified the device used by participants at the beginning of the survey. Respondents attempting to access the web survey with a non-smartphone device were unable to proceed and were instructed to use a smartphone. At the start of the smartphone survey, respondents were randomly assigned to one of two experimental conditions. Respondents in the written condition received two open-ended probing questions (OPQs) with a request for written answers, while respondents in the oral condition received the same probing questions but with a request for oral answers.

Importantly, this study was part of a larger smartphone survey with several studies concerning written and oral answers (see, e.g., Höhne et al. 2024). The study was placed in the center of the smartphone survey. At this point, respondents had already been asked six open-ended questions with requests for written and oral answers.

3.2 Sample

Forsa invited 6,745 respondents to participate in the web survey. None of the respondents were excluded because of full quotas or attempting to access the web survey with a device other than a smartphone. Out of the 1,681 respondents who started the web survey, 680 broke off before answering any study-related questions. In the written condition, 159 (approximately 24 percent) respondents broke off, whereas in the oral condition, 521 (approximately 51 percent) respondents broke off.

There were 1,001 respondents who completed the survey, with 500 in the written condition and 501 in the oral condition. The AAPOR Response Rate 1 was about 15 percent (American Association for Public Opinion Research 2023). Table 1 shows the sample characteristics of the written and oral conditions. In order to evaluate the effectiveness of random assignment, we compared the sample composition between the two conditions. We found no

Table 1. Sample Characteristics by Experimental Condition

Respondent characteristics	Written condition	Oral condition
Age	48.1 (15.1)	48.7 (14.5)
Gender: female	50.0	48.3
Education: medium	41.0	42.5
Education: high	29.2	26.4
Smartphone skills	5.6 (1.2)	5.6 (1.2)
Internet usage via smartphone	6.1 (1.2)	6.0 (1.4)

NOTE.—$n = 500$ (written) and $n = 501$ (oral). We report means and standard deviations (in parentheses) for age, smartphone skills, and Internet usage via smartphone. For gender and education, we report percentages. There were no missing data on any of these variables. English translations of the questions used to collect the respondent characteristics are listed in Supplementary Appendix D.

statistically significant differences with respect to age, gender, education, smartphone skills, and Internet usage.

3.3 Survey and Open-Ended Probing Questions

We asked two survey questions (SQs) regarding citizens' relationship with the state, which were adopted from the National Identity and Citizenship modules of the German questionnaires of the International Social Survey Programme (ISSP) 2013 and 2014. Each survey question was followed by an open-ended probing question (OPQ), requesting either written or oral answers. Figure 1 provides exemplary screenshots of the survey and open-ended probing questions.

The following formulations are English translations of the original German wordings, which are documented in Supplementary Appendix A (including the response distributions of both SQs).

> SQ1: To what extent do you agree or disagree with the following statement? I feel more like a citizen of the world, and thus connected to the world as a whole, and less as a citizen of a particular country. Response categories: 1 "Strongly agree," 2 "Agree," 3 "Neither nor," 4 "Do not agree," 5 "Do not agree at all," and 6 "Can't say."
>
> OPQ1: How did you understand the term 'citizen of the world' in the last question? "Please enter your answer into the text field." or "Press and hold the microphone icon while recording your answer."
>
> SQ2: There are different views about people's rights in a democracy. How important is it that citizens may engage in acts of civil disobedience when they strictly oppose government actions? Response categories: 1 "Not at all important" to 7 "Very important" and 8 "Can't say."

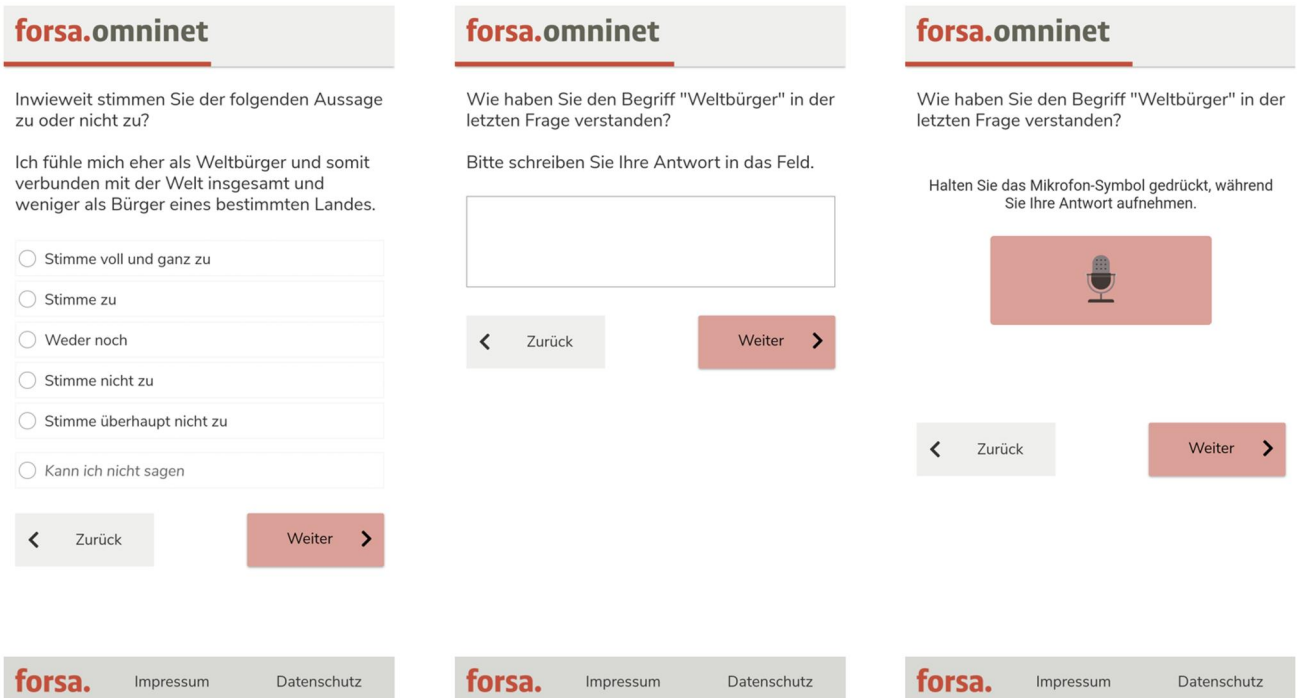

Figure 1. Exemplary Screenshots of the Survey and Open-Ended Probing Questions. Survey question on "citizen of the world" on the left, open-ended probing question with a request for a written answer in the middle, and open-ended probing question with a request for an oral answer on the right.

OPQ2: How did you understand the term "civil disobedience" in the last question? Please provide examples. "Please enter your answer into the text field." or "Press and hold the microphone icon while recording your answer."

The rating scales of the two survey questions were vertically aligned with non-substantive categories visually separated. At the survey start, respondents received a description on how to provide written or oral answers, depending on the experimental condition. Supplementary Appendix B provides English translations of these descriptions.

In order to capture respondents' oral answers, we integrated the open-source tool SurveyVoice (SVoice; Höhne et al. 2021) into the Forsa web survey system. SVoice employs various programming languages, such as JavaScript and HTML, to record oral answers via smartphones' built-in microphones. SVoice works on both Android and iOS smartphones.

3.4 Coding Procedure

Prior to analysis, the audio recordings were transcribed by a student assistant, who was instructed to transcribe the content verbatim, but to omit hesitation markers and fillers, such as "um" or "uh." These hesitation markers and fillers did not require removal from the written responses, as they were absent. As a quality assurance measure, 10 percent of the transcripts ($n = 62$) were double-checked by the first author, requiring only minor corrections, such as spelling errors.

Answers to the two open-ended probes were coded into *themes* and *theme areas* by the first author using the constant comparative method (Ridolfo and Schoua-Glusberg 2011). We used an inductive coding approach and developed the themes and theme areas based on the data rather than using preconceived codes. Theme areas are larger categories that include several similar, yet distinctive themes (e.g., the themes "non-violent protests," "express opinion publicly," and "disruptive, but peaceful behavior" all belong to the theme area "peaceful, active forms of protest;" see table 4). The final coding schemes comprised six (OPQ1) and five (OPQ2) *theme areas*, including 18 (OPQ1) and 12 (OPQ2) *themes*, respectively (see tables 3 and 4).

Furthermore, we used four additional codes for responses that could not be evaluated in terms of content, two for "probe nonresponse" and two for "not clearly interpretable answers." Similar to Meitinger and Kunz (2022), we define *probe nonresponse* as instances in which respondents gave no answer at all (i.e., left the text field blank or did not record a voice answer; *complete nonresponse*) and instances in which respondents implicitly refused to answer by providing useless answers (e.g., "yes," "nonsense," "stupid question"; *soft nonresponse*). In contrast to Meitinger and Kunz (2022), however, we do not distinguish between the two forms of nonresponse in our analyses for the following reasons: first, the distinction between complete and soft nonresponse is not of practical relevance since neither form of "answer" can be evaluated in

terms of content analysis. Second, the soft nonresponse rate was very low (OPQ1: $n = 35$; OPQ2: $n = 29$) impeding the conduction of robust statistical analyses. *Not clearly interpretable answers* were defined as answers that either did not match the probing question (e.g., when respondents elaborated on why they selected their answer instead of explaining what they understood by the term civil disobedience in OPQ2) or were uninterpretable in the context of the probe (e.g., "A global person" in OPQ1 or "If you disagree with the government" in OPQ2).

Using the final coding schemes—with the additional three codes for "soft probe nonresponse," "probe-answer mismatch," and "uninterpretable answers"—a student assistant independently coded a randomly selected subset of 25 percent (OPQ1: $n = 200$, OPQ2: $n = 191$) of the probe answers, so that we could estimate interrater reliability. The size of the subset of probe answers and number of coders were chosen for time and budget reasons. Interrater agreement was calculated based on the *theme areas* and deemed substantial (Landis and Koch 1977) with Cohen's kappa values of 0.79 for both questions (agreement rate: OPQ1 = 83.5 percent and OPQ2 = 85.3 percent). The interrater agreement and kappa values indicate that the number of coders was adequate (see Nunnally (1978) and Rust and Cooil (1994) for a comprehensive discussion of minimally acceptable levels of interrater agreement and appropriate numbers of judges). The two coders discussed any discrepancies between the two ratings until consensus was reached.

3.5 Analytical Strategies

Data preparation and analyses were conducted with Stata 18 and Microsoft Excel 365. To examine our first research question, we compare the probe answers in the two experimental conditions regarding the following data quality indicators: probe nonresponse (hypothesis 1), not clearly interpretable answers (hypothesis 2), and word count (hypothesis 3). As mentioned above, *probe nonresponse* includes instances of complete as well as soft nonresponse and *not clearly interpretable answers* includes answers that either do not match the probing question or are uninterpretable in the context of the probe. *Word count* refers to the number of words respondents provided in their answers to the two probes, respectively. Word count was determined in Microsoft Excel 365 using the formula "=LEN(Cell) – LEN (SUBSTITUTE (Cell; "";"")) + 1)." Importantly, the probe nonresponse analyses were based on the whole sample ($N = 1{,}001$). Answers coded as nonresponse (complete or soft) were excluded from the analyses of not clearly interpretable answers and word count, leaving 765 (OPQ1) and 736 (OPQ2) cases in these analyses, respectively. Supplementary Appendix C reports the case numbers for the coded metrics.

In a first step, we perform directional Z-tests for investigating hypotheses 1 and 2 on probe nonresponse and not clearly interpretable answers. For

investigating hypothesis 3 on word count, we conduct nonparametric Mann–Whitney *U*-tests to account for the non-normally distributed data. In a second step, we estimate logistic regression models with probe nonresponse (1 = yes) and not clearly interpretable answers (1 = yes) as dependent variables for each of the two OPQs. We also estimate negative binomial regression models with word count as dependent variable for each of the two OPQs. In all regression models, we include oral communication mode (1 = yes) as the main independent variable and additionally control for respondents' gender (1 = female), age (in years), and educational attainment (in the form of two dummy variables with low as a reference: medium = 1 and high = 1). In doing so, we examine whether and to what extent effects of the communication mode were affected by respondent characteristics that potentially impact respondent behavior (see van Vaerenbergh and Thomas 2013). English translations of all variables are documented in Supplementary Appendix D.

In the analyses of our second research question, we first look at differences in the average number of themes mentioned between the two conditions (hypothesis 4) using Mann–Whitney *U*-tests. In a second step, we estimate zero-truncated Poisson regression models for each of the two OPQs with number of themes mentioned as dependent variables. Again, we include oral communication mode as the main independent variable and additionally control for respondents' gender, age, and educational attainment. Next, we compare the overlap of the themes being mentioned in the two conditions (hypothesis 5). Following Meitinger and Behr (2016) and Gavras et al. (2022), we analyze the overlap of themes descriptively. These analyses are constrained to respondents who gave interpretable answers to the two open-ended probes (OPQ1: $n = 652$, OPQ2: $n = 666$).

For examining our third research question, we compare the two conditions with respect to respondents' *survey evaluations* at the end of the survey (hypothesis 6): the extent to which respondents rated the survey as interesting (1 = not at all to 7 = very), difficult (1 = very easy to 7 = very difficult), and lengthy (1 = not at all to 7 = very). These analyses are based on the whole sample ($N = 1{,}001$). We first conduct Mann–Whitney *U*-tests, followed by estimating three ordinal logistic regression models with survey evaluations (coded from 1 to 7, respectively) as dependent variables. Similar to the regression analyses reported above, we include oral communication mode as the main independent variable and additionally control for respondents' gender, age, and educational attainment.

4. RESULTS

4.1 Research Question 1

As shown in table 2, probe nonresponse was about four to five times higher in the oral than the written condition. This similarly applies to both probes. The

Table 2. Statistics of the Data Quality Indicators by Experimental Condition

Data quality indicators	Written			Oral			Test	
	%		*n*	%		*n*	*Z*	*p*
Probe nonresponse								
OPQ1	8.2		41	38.9		195	–11.45	.000
OPQ2	11.2		56	41.7		209	–10.94	.000
Not clearly interpretable answers								
OPQ1	17.2		79	11.1		34	2.33	.010
OPQ2	10.4		46	8.2		24	0.97	.166
	M (SD)	Mdn	*n*	*M* (SD)	Mdn	*n*	*z*	*p*
Word count								
OPQ1	14.18 (12.19)	11.0	459	43.82 (36.93)	32.0	306	–15.38	.000
OPQ2	12.08 (11.71)	9.0	444	41.19 (38.41)	29.0	292	–14.66	.000

NOTE.—*M* = mean; SD = standard deviation; Mdn = median.

results of directional Z-tests were statistically significant, providing supportive evidence for our first hypothesis. However, we only found partial support for our second hypothesis. As shown in table 2, the share of not clearly interpretable answers was significantly (and about 6 percentage points) lower in the oral than the written condition for OPQ1. For OPQ2, we observe a similar tendency. However, the result of the directional Z-test was not significant. The results of word count support our third hypothesis: respondents in the oral condition used about three times as many words to answer both probes than respondents in the written condition. The results of two Mann–Whitney *U*-tests were statistically significant, indicating that requests for oral answers indeed trigger open narrations.

In order to test the robustness of our results, we estimated logistic (nonresponse and not interpretable answers) and negative binomial regression models (word count) controlling for respondents' age, gender, and education. The overall conclusions did not change indicating the robustness of our results (see tables E1 to E3 in Supplementary Appendix E).

4.2 Research Question 2

To examine our second research question, we first analyzed the average number of themes mentioned in both conditions. Respondents mentioned between one and three themes in their answers to the two probing questions. Overall, respondents in the oral condition mentioned more themes in response to OPQ1 ($M = 1.38$, $SD = 0.59$, $Mdn = 1.0$, $n = 272$) than participants in the written

condition ($M = 1.13$, $SD = 0.38$, $Mdn = 1.0$, $n = 380$). This difference was small, but statistically significant ($z = –6.36$, $p = .000$). The average number of themes mentioned in response to OPQ2 did not differ between both conditions (oral: $M = 1.28$, $SD = 0.50$, $Mdn = 1.0$, $n = 268$; written: $M = 1.24$, $SD = 0.47$, $Mdn = 1.0$, $n = 398$; $z = –0.94$, $p = .346$). Hence, there is only partial evidence for our fourth hypothesis. We tested the robustness of our results by estimating two zero-truncated Poisson regression models controlling for age, gender, and education. Again, the results remained unchanged (see table E4 in Supplementary Appendix E).

Next, we examined differences in the themes mentioned between the two conditions. The results are displayed in table 3 (OPQ1) and table 4 (OPQ2). The number and types of *theme areas* that respondents mentioned in response to the two open-ended probes were very similar in both conditions. We only found small differences in the frequency with which some *themes* were mentioned. However, these differences were relatively small and primarily appeared in topics that were mentioned by only a few respondents. Except for two themes that were only mentioned by one or two respondents in the written condition (OPQ1: "Sb. who speaks many languages," OPQ2: "Right of resistance under the German Basic Law, Art. 20, Par. 4 GG"), all themes were mentioned by at least one respondent in each condition. All in all, we found no descriptive support for our fifth hypothesis.

In response to OPQ1, respondents most frequently explained their understanding of the term citizen of the world by referring to a person's (perceived) citizenship (written: 31.1 percent, oral: 32.7 percent). The second most frequently mentioned theme area was that the term refers to a person with an open-minded attitude (written: 33.4 percent, oral: 29.4 percent). Several respondents in both conditions explained that they did not understand the term (e.g., "I have no idea what the word means.") or showed an incorrect understanding of it (e.g., "Inhabitant of planet Earth").

When probed for their understanding of the term civil disobedience in OPQ2, respondents most frequently mentioned peaceful, active forms of protest (written: 57.0 percent, oral: 55.2 percent). This was followed by peaceful, passive forms of protest (written: 23.1 percent, oral: 19.8 percent). Several respondents in both conditions explained that they did not understand the term (e.g., "Unfortunately, I did not understand this term at all.") and others misinterpreted it (e.g., "holding referendums").

4.3 Research Question 3

Turning to our sixth research hypothesis, we found no differences between the two conditions in how interesting or difficult respondents evaluated the survey. However, respondents in the oral condition rated the survey as significantly shorter than respondents in the written condition (see table 5). Given the fact

Table 3. Percentages (Frequencies) and Number of Theme Areas and Themes Mentioned in Response to OPQ1 (Citizen of the World) by Experimental Condition

Theme areas/themes	Written		Oral	
	%	*n*	%	*n*
Citizenship/belonging to a country	31.1	118	32.7	89
Sb. who does not feel that they belong to any particular country (but to the world as a whole).	24.5	93	31.3	85
Sb. who denies their homeland/has no national pride	2.6	10	0.4	1
Sb. who is not patriotic	1.8	7	0.4	1
Sb. who has no citizenship/has discarded it	1.6	6	0.4	1
Sb. who has multiple citizenships	0.5	2	0.4	1
Open-minded attitude	33.4	127	29.4	80
Sb. who thinks globally	10.5	40	12.1	33
Sb. who is open-minded toward others/new things	10.8	41	8.5	23
Sb. for whom a person's origin is irrelevant, who sees all people as equal	9.5	36	7.7	21
Sb. who is tolerant/free of prejudice	2.6	10	1.1	3
Not bound to a specific place of residence	17.6	67	19.5	53
Sb. who feels at home all over the world	14.5	55	15.4	42
Sb. who travels a lot	2.4	9	3.3	9
Sb. who lives/has lived in different countries	0.5	2	0.7	2
Sb. who speaks many languages	0.3	1	0	0
Responsibility for the whole world/planet Earth	7.6	29	8.5	23
Sb. who feels responsible for the whole world	3.9	15	6.6	18
Sb. who bears responsibility for planet Earth and his/her actions (e.g., consumption, traveling)	3.7	14	1.8	5
Term is criticized	4.7	18	7.0	19
Respondent does not understand the term	3.4	13	4.8	13
Respondent rejects the term	1.3	5	2.2	6
Inhabitant of planet Earth	5.5	21	2.9	8

NOTE.—$n = 380$ (written) and $n = 272$ (oral). Sb. = somebody.

Table 4. Percentages (Frequencies) and Number of Theme Areas and Themes Mentioned in Response to OPQ2 (Civil Disobedience) by Experimental Condition

Theme areas/themes	Written		Oral	
	%	*n*	%	*n*
Peaceful, active forms of protest	57.0	227	55.2	148
Non-violent protests (e.g., demonstrations, strikes)	42.0	167	40.7	109
Express opinion publicly	12.1	48	11.2	30
Disruptive, but peaceful behavior (generates costs, impairs third parties)	3.0	12	3.4	9
Peaceful, passive forms of protest	23.1	92	19.8	53
Non-compliance with government decisions/recommendations (e.g., vaccination, masking)	16.1	64	13.8	37
Disobeying laws/rules (non-violent behavior)	7.0	28	6.0	16
Term is criticized	8.0	32	11.2	30
Respondent does not understand the term	6.0	24	9.3	25
Respondent misinterprets the term	1.3	5	1.5	4
Respondent rejects the term	0.8	3	0.4	1
Ambiguous forms of protest	7.8	31	8.2	22
Violation of law/crime (unclear whether non-violent or not)	3.3	13	4.5	12
Revolt/protest against the state (unclear whether non-violent or not)	4.0	16	3.7	10
Right of resistance under the German Basic Law, Art. 20, Par. 4 GG	0.5	2	0	0
Violent, active forms of protest (e.g., damaging property, injuring people)	4.0	16	5.6	15

NOTE.—$n = 398$ (written) and $n = 268$ (oral).

that probe nonresponse was higher in the oral than in the written condition, one could argue that the difference in perceived survey length is because more respondents provided no answer to the open-ended probes in the oral condition. We therefore re-ran the analysis excluding nonrespondents, but the results did not change (written condition: $M = 3.41$, $SD = 1.62$, $Mdn = 4.0$, $n = 456$; oral condition: $M = 2.59$, $SD = 1.40$, $Mdn = 3.0$, $n = 282$; $z = 6.79$, $p = .000$). Finally, we tested the robustness of our results by estimating ordinal logistic regression models using age, gender, and education as controls. The results remained unchanged (see table E5 in Supplementary Appendix E).

Table 5. Statistics of the Survey Evaluation Questions by Experimental Condition

Survey evaluation questions	Written			Oral			Test	
	M	SD	Mdn	*M*	SD	Mdn	*z*	*p*
Interest	5.43	1.36	6.0	5.38	1.57	6.0	–0.34	.731
Difficulty	2.94	1.42	3.0	3.10	1.52	3.0	–1.58	.115
Length	3.36	1.63	3.0	2.64	1.43	3.0	7.10	.000

NOTE.—M = mean; SD = standard deviation; Mdn = median. $n = 500$ (written) and $n = 501$ (oral).

5. DISCUSSION AND CONCLUSION

This study aimed to explore the quality of answers obtained when asking respondents to provide oral answers to open-ended probing questions in a web probing study. To this end, we set up an experiment in a smartphone survey and examined differences between the answers to two comprehension probes given by respondents who were asked to provide written answers and those who were asked to provide oral answers recorded through the built-in microphone of their smartphone.

Three research questions guided our analysis, each accompanied by tailored hypotheses. First, we examined the data quality of respondents' probe answers in terms of probe nonresponse, not clearly interpretable answers, and word count. As expected, respondents were more reluctant to provide oral answers compared to written ones. However, oral answers were significantly longer, and for one of the two probes (OPQ1), they were more clearly interpretable (and thus analyzable) than written answers. This suggests that while many respondents may not fully embrace this new methodology, those who do, take up its benefits and give more extensive answers. Potentially, these longer answers can help to reduce the proportion of uninterpretable answers in web probing studies.

Second, we examined whether the two modes of communication influenced the content of respondents' probe answers. For the first of the two comprehension probes, we found the expected difference in the number of themes mentioned by respondents, which was higher in the oral than in the written condition. However, when comparing the types of themes mentioned by respondents between the two conditions, we only found relatively small differences. This latter finding suggests that the open narrations triggered by requests for oral answers do not lead to more detailed or thoughtful answers. On the contrary, one could argue that requests for written answers lead respondents to provide the same information in terms of content, but in a much more concise manner than requests for oral answers.

Our results contradict findings reported by Gavras et al. (2022), who found only little overlap between themes mentioned in written and oral answers in a smartphone survey. One explanation for the disagreement of these results could be that in the latter study, respondents were asked to answer open-ended survey questions, while we asked them to answer open-ended probes about their understanding of specific terms in previous survey questions. These probes are much more targeted than open-ended survey questions and the universe of possible interpretations of a term is limited, so that differences in the themes being mentioned are rather unlikely with larger sample sizes. Of course, this is only an *ad hoc* explanation that requires additional research. Another point is that Gavras et al. (2022), in contrast to us, employed text-as-data-methods in the form of Structural Topic Models (Roberts et al. 2014). Thus, it seems worthwhile to compare the alignment of human and automated text coding strategies in future studies.

The results on number of words and number of themes may be related to an interplay between the more intuitive and spontaneous online processing and a less burdensome answer delivery associated with oral answers. To put it differently, respondents only need to press a recording button and record their answer while drawing on a preconfigured online-tally in their mind. If respondents are asked about their understanding of "civil disobedience" (OPQ2), they may not search for relevant information in their memory but refer to a tally (i.e., a summative impression of the object under investigation; Kim and Garrett 2012) that includes previously stored aspects related to this topic (see Lodge et al. 1989; McGraw et al. 2003). This is only an attempted explanation that lacks solid empirical evidence. Thus, we recommend that future research includes fine-grained measures of answer delivery burden (e.g., response times and self-reports), including techniques for unraveling response processes, such as think aloud. This may help to draw more robust conclusions when it comes to written and oral answers in the web survey context.

Our third research question concerned respondents' evaluation of the survey with respect to interest, difficulty, and length. Ratings of interest and difficulty did not substantially differ between the two conditions, suggesting that neither communication mode promotes satisficing answer behavior to a greater extent. Particularly, the finding on difficulty indicates that both answer formats do not result in different levels of respondent fatigue. It is important to note that the Forsa Omninet Panel—similar to most other online panels—uses text-based answer formats as default. Respondents are not used to requests for oral answers so that the lack of familiarity may have increased their perceived difficulty. It would be worthwhile to compare requests for oral and written answers in a freshly recruited panel or sample. This way, it would be possible to better disentangle the interplay between perceived difficulty and answer format standard. Looking at length, however, revealed that respondents in the oral condition rated the survey as shorter, suggesting a higher perceived time-efficiency of oral answers. Unfortunately, response times were not collected in

this smartphone survey, so we do not know whether respondents actually required less time to complete the survey or whether they only perceived it to be less time-consuming. Future research could investigate whether completion times differ between the two communication modes.

Overall, our results present a mixed picture, with no clear preference for one communication mode over the other. While requests for oral answers yielded longer and more interpretable answers, they also resulted in higher probe nonresponse. The types of themes mentioned by respondents did not differ between both communication modes, suggesting similar data quality from a content perspective. Regarding the practical implications of our results for web probing studies, we must conclude that there is yet little evidence that justifies a shift from the written to the oral communication mode when it comes to comprehension probes. Since oral answers need to be transcribed before analysis, this appears even less reasonable at the moment. However, the emergence of speech-to-text APIs, such as OpenAI's Whisper (Radford et al. 2023), may offer a more time- and cost-efficient way of transcribing oral answers from smartphone surveys in the near future.

Like most empirical studies, our study has some methodological limitations that provide avenues for future research. First, we used only two probing questions of the same type (comprehension probes), which may limit the generalizability of our results. It is conceivable that other probe types, such as category-selection probes that ask respondents to give reasons for their answer to an survey question, trigger different response processes leading to (larger) differences between the two communication modes. Hence, it would be fruitful to extend the methodological investigation to further questions and probe types. Second, respondents were drawn from a non-probability online panel. Given that most web probing studies make use of non-probability samples, this does not necessarily restrict the generalizability to typical web probing studies, but certainly to web probes implemented in surveys with probability samples. Going beyond non-probability samples would have the advantage to draw more robust and general conclusions. Third, respondents were randomly assigned to oral or written request conditions. However, it would also be possible to let respondents decide for themselves which communication mode to participate in a web probing study. This may decrease probe nonresponse and attract respondent groups, such as young people, that frequently hesitate to take part in conventional web surveys (Revilla and Höhne 2020). Relatedly, it might be worthwhile to include a more sophisticated incentive strategy. For example, it would be possible to incentivize respondents per (oral or written) answer instead of a fixed, overall incentive. Finally, our experiment was implemented in a smartphone survey. As mentioned above, we know from previous research that many respondents dislike the thought of answering surveys on their smartphones but prefer completing them on a PC. Giving respondents the possibility to record their answers on a PC or laptop may also help to increase their willingness to answer open-ended probes with requests for oral answers.

Supplementary Materials

Supplementary materials are available online at academic.oup.com/jssam.

REFERENCES

American Association for Public Opinion Research (2023), Standard Definitions: Final Disposition of Case Codes and Outcome Rates for Surveys (10th ed.), Alexandria, VA: AAPOR.

Behr, D., Kaczmirek, L., Bandilla, W., and 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.

Collins, D. (ed.) (2015), *Cognitive Interviewing Practice*, London: Sage Publications.

Conrad, F. G., and Blair, J. (2009), "Sources of Error in Cognitive Interviews," *Public Opinion Quarterly*, 73, 32–55.

Converse, J. M., and Presser, S. (1986), *Survey Questions: Handcrafting the Standardized Questionnaire*, Beverly Hills, CA: Sage.

Fowler, S. L., and Willis, G. B. (2020), "The Practice of Cognitive Interviewing Through Web Probing," in *Advances in Questionnaire Design, Development, Evaluation and Testing*, eds. P. Beatty, D. Collins, L. Kaye, J. L. Padilla, G. B. Willis, and A. Wilmot, Hoboken, NJ: Wiley, pp. 451–469.

Gavras, K., and Höhne, J. K. (2020), "Evaluating Political Parties: Criterion Validity of Open Questions With Requests for Text and Voice Answers," *International Journal of Social Research Methodology*, 25, 135–141.

Gavras, K., Höhne, J. K., Blom, A., and Schoen, H. (2022), "Innovating the Collection of Open-Ended Answers: The Linguistic and Content Characteristics of Written and Oral Answers to Political Attitude Questions," *Journal of the Royal Statistical Society (Series A)*, 185, 872–890.

Gummer, T., Höhne, J. K., Rettig, T., Roßmann, J., and Kummerow, M. (2023), "Is There a Growing Use of Mobile Devices in Web Surveys? Evidence From 128 Web Surveys in Germany," *Quality & Quantity*, 57, 5333–5353.

Hadler, P., Neuert, C., Lenzner, T., and Menold, N. (2018), "European Working Conditions Survey (EWCS) (English Version): Cognitive Pretest," GESIS—Leibniz Institute for the Social Sciences, GESIS Project Reports. Available at https://doi.org/10.17173/pretest72.

Höhne, J. K., Gavras, K., and Claassen, J. (2024), "Typing or Speaking? Comparing Text and Voice Answers to Open Questions on Sensitive Topics in Smartphone Surveys," *Social Science Computer Review*, Available at https://doi.org/10.1177/08944393231160961.

Höhne, J. K. (2023), "Are Respondents Ready for Audio and Voice Communication Channels in Online Surveys?" *International Journal of Social Research Methodology*, 26, 335–342.

Höhne, J. K., Gavras, K., and Qureshi, D. D. (2021), "SurveyVoice (SVoice): A Comprehensive Guide for Collecting Voice Answers in Surveys," *Zenodo*. Available at https://doi.org/10.5281/zenodo.4644590.

Kim, Y. M., and Garrett, K. (2012), "On-Line and Memory-Based: Revisiting the Relationship Between Candidate Evaluation Processing Models," *Political Behavior*, 34, 345–368.

Krosnick, J. A. (1991), "Response Strategies for Coping with the Cognitive Demands of Attitude Measures in Surveys," *Applied Cognitive Psychology*, 5, 213–236.

Landis, R. J., and Koch, G. G. (1977), "The Measurement of Observer Agreement for Categorical Data," *Biometrics*, 33, 159–174.

Lenzner, T., and Höhne, J. K. (2022), "Who Is Willing to Use Audio and Voice Inputs in Smartphone Surveys, and Why?" *International Journal of Market Research*, 64, 594–610.

Lenzner, T., and Neuert, C. (2017), "Pretesting Survey Questions via Web Probing—Does It Produce Similar Results to Face-to-Face Cognitive Interviewing?," *Survey Practice*, 10, 1.

Lenzner, T., Schick, L., Hadler, P., Behnert, J., Steins, P., and Neuert, C. (2022), "FGZ Cohesion Panel: Wave 2—Questions on Climate Change, Antisemitism, and Gender Equality (English

Version). Cognitive Online Pretest," GESIS—Leibniz Institute for the Social Sciences, GESIS Project Reports. Available at https://doi.org/10.17173/pretest129.
Lodge, M., McGraw, K. M., and Stroh, P. (1989), "An Impression-Driven Model of Candidate Evaluation," *American Political Science Review*, 83, 399–419.
Lugtig, P., and Toepoel, V. (2016), "The Use of PCs, Smartphones, and Tablets in a Probability-Based Panel Survey: Effects on Survey Measurement Error," *Social Science Computer Review*, 34, 78–94.
McGraw, K. M., Hasecke, E., and Conger, K. (2003), "Ambivalence, Uncertainty, and Processes of Candidate Evaluation," *Political Psychology*, 24, 421–448.
Meitinger, K. (2017), "Necessary But Insufficient: Why Measurement Invariance Tests Need Online Probing as a Complementary Tool," *Public Opinion Quarterly*, 81, 447–472.
Meitinger, K., and Behr, D. (2016), "Comparing Cognitive Interviewing and Online Probing: Do They Find Similar Results?" *Field Methods*, 28, 363–380.
Meitinger, K., and Kunz, T. (2022), "Visual Design and Cognition in List-Style Open-Ended Questions in Web Probing," *Sociological Methods & Research*, 53, 940–967.
Miller, K., Willson, S., Chepp, V., and Padilla J.-L. (eds.) (2014), *Cognitive Interviewing Methodology*, Hoboken, NJ: Wiley.
Neuert, C. E., Meitinger, K., and Behr, D. (2021), "Open-Ended Versus Closed Probes: Assessing Different Formats of Web Probing," *Sociological Methods & Research*, 52, 1981–2015.
Nunnally, J. C. (1978), *Psychometric Theory*, New York, NY: McGraw-Hill.
Padilla, J.-L., and Benítez, I. (2014), "Validity Evidence Based on Response Processes," *Psicothema*, 26, 136–144.
Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., and Sutskever, I. (2023), "Robust Speech Recognition via Large-Scale Weak Supervision," in Proceedings of the 40th International Conference on Machine Learning, Honolulu, Hawaii, USA, pp. 28492–28518. Available at https://dl.acm.org/doi/10.5555/3618408.3619590
Revilla, M., and Couper, M. P. (2021), "Improving the Use of Voice Recording in a Smartphone Survey," *Social Science Computer Review*, 39, 1159–1178.
Revilla, M., Couper, M. P., Bosch, O. J., and Asensio, M. (2020), "Testing the Use of Voice Input in a Smartphone Web Survey," *Social Science Computer Review*, 38, 207–224.
Revilla, M., and Höhne, J. K. (2020), "Comparing the Participation of Millenials and Older Age Cohorts in the CROss-National Online Survey Panel and the German Internet Panel," *Survey Research Methods*, 14, 499–513.
Revilla, M., and Ochoa, C. (2016), "Open Narrative Questions in PC and Smartphones: Is the Device Playing a Role?," *Quality & Quantity*, 50, 2495–2513.
Revilla, M., Toninelli, D., Ochoa, C., and Loewe, G. (2016), "Do Online Access Panels Really Need to Allow and Adapt Surveys to Mobile Devices?" *Internet Research*, 26, 1209–1227.
Ridolfo, H., and Schoua-Glusberg, A. (2011), "Analyzing Cognitive Interview Data Using the Constant Comparative Method of Analysis to Understand Cross-Cultural Patterns in Survey Data," *Field Methods*, 23, 420–438.
Roberts, M. E., Stewart, B. M., Tingley, D., Lucas, C., Leder-Luis, J., Kushner Gadarian, S., Albertson, B., and Rand, D. G. (2014), "Structural Topic Models for Open-Ended Survey Responses," *American Journal of Political Science*, 58, 1064–1082.
Rust, R. T., and Cooil, B. (1994), "Reliability Measures for Qualitative Data: Theory and Implications," *Journal of Marketing Research*, 31, 1–14.
Schick, L., Lenzner, T., Hadler, P., and Neuert, C. (2023), "FReDA-W3b—Fragen Zu Den Themen Partnerschaftsstatus, Ernährungsstile, Globale Unsicherheit Und Vertrauen in Institutionen. Kognitiver Online-Pretest," GESIS—Leibniz Institute for the Social Sciences, GESIS Project Reports, Available at https://doi.org/10.17173/pretest127.
Schober, M. F., Conrad, F. G., Antoun, C., Ehlen, P., Fail, S., Hupp, A. L., Johnston, M., Vickers, L., Yan, H. Y., and Zhang, C. (2015), "Precision and Disclosure in Text and Voice Interviews on Smartphones," *PloS One*, 10, e0128337.
Struminskaya, B., Weyandt, K., and Bosnjak, M. (2015), "The Effects of Questionnaire Completion Using Mobile Devices on Data Quality. Evidence From a Probability-Based General Population Panel," *Methods, Data, Analyses*, 9, 261–292.

Tourangeau, R., Rips, L. J. and Rasinski, K. A. (2000), *The Psychology of Survey Response*, Cambridge: Cambridge University Press.
van Vaerenbergh, Y., and Thomas, T. D. (2013), "Response Styles in Survey Research: A Literature Review of Antecedents, Consequences, and Remedies," *International Journal of Public Opinion Research*, 25, 195–217.
Willis, G. B. (2005), *Cognitive Interviewing: A Tool for Improving Questionnaire Design*, Thousand Oaks: Sage.
Willis, G. B., and Miller, K. (2011), "Cross-Cultural Cognitive Interviewing: Seeking Comparability and Enhancing Understanding," *Field Methods*, 23, 331–341.
Zaller, J., and Feldman, S. (1992), "A Simple Theory of the Survey Response: Answering Questions versus Revealing Preferences," *American Journal of Political Science*, 36, 579–616.

Journal of Survey Statistics and Methodology, 2024, 00, 1–23
https://doi.org/10.1093/jssam/smae031
Article