Response Burden and Response Quality in Web Probing. An Experiment on the Effects of Probe Placement and Format

Patricia Hadler[1]

[1]GESIS - Leibniz Institute for the Social Sciences

Probes are follow-up questions to survey questions used for question pretesting and evaluation. A probe can be placed concurrently - that is directly following the survey question it relates to - or retrospectively later or at the end of a survey. Retrospective probing ensures that the natural flow of the survey questions is not interrupted. In contrast, concurrent probing is argued to prioritize the quality of probe responses, as respondents can still access their short-term memory. To date, concerns that retrospective probe placement negatively impacts how probes are perceived and answered have not been tested experimentally. Recently, it has been suggested to employ "closed" probes with predefined response options, for instance using a check-all-that-apply format, also with the rationale of improving probe response quality. Closed probe formats suffer from fewer nonresponse and more themes being chosen than being typed in response to open-ended probes.

The present research reports a study on the impact of probe placement on response burden and response quality of web probes. Moreover, it examines whether the effects of probe placement are moderated by probe format. Results show that respondents needed longer to recapitulate the survey question when probes are asked retrospectively and were less likely to offer interpretable probe response content for one of three examined probes. These effects were not moderated by probe format. However, respondents who answered open-ended probes retrospectively were more likely to rely on contextual cues from intermittent survey questions. This effect did not occur when respondents were presented probes with predefined response options. Implications for web probing design are discussed.

Keywords: web surveys; context effects; question order effects; cognitive pretesting; open-ended questions

1   Introduction

Cognitive pretesting is a method of question evaluation in which respondents reflect on survey questions and their answers to them (Beatty & Willis, 2007; Presser et al., 2004).

Corresponding author: Patricia Hadler, GESIS - Leibniz Institute for the Social Sciences, 68159 Mannheim, Germany (Email: patricia.hadler@gesis.org)

It examines how respondents construct the pragmatic meaning of a survey question (Miller et al., 2014) and seeks to identify problems respondents encounter during survey response, such as comprehension issues or choosing a suitable response category (Tourangeau et al., 2000). These insights can be used to revise the questions and increase survey data quality (Lenzner, Neuert, & Otto, 2016).

Cognitive pretesting has traditionally been carried out in the form of face-to-face interviews (Collins, 2015; Willis, 2005), in which interviewers may employ the technique of asking *probes*, that is questions about the survey question, such as how respondents understood a particular term or why they chose a specific answer category (Foddy, 1998). Web probing implements techniques from cognitive interviewing into (self-administered) web surveys (Behr,

Kaczmirek, Bandilla, & Braun, 2012a; Edgar et al., 2016; Meitinger & Behr, 2016). The benefits of web probing include the possibility of collecting data from large sample sizes quickly (Meitinger & Behr, 2016) while avoiding the labour-intensive transcribing of personal interviews (Willis, 2015a). Moreover, web probes can be implemented into production surveys to support the interpretation of survey findings (Meitinger, 2017; Singer & Couper, 2017).

A fundamental research design decision when implementing probes is probe *placement*, or when to ask the probing question (Willis, 2005, p. 51 f.). One possibility in web probing is to embed the probe alongside the survey question on the same survey page (i.e., Couper, 2013; Luebker, 2021). More common, however, is to disentangle the response process of the survey question from the probing process (Behr et al., 2017; Converse & Presser, 1986) by either placing the probe *concurrently*, that is, directly following the survey question but on a separate page, or *retrospectively* after a block of survey questions or even at the end of a questionnaire (Collins, 2015, p. 120). The rationale behind concurrent probing is to ensure that respondents' thought processes are still available in short-term memory. Retrospective probing is implemented so as not to interrupt the flow of a questionnaire and to prevent probes from interfering with subsequent survey questions. In a nutshell, concurrent probing is argued to prioritize the quality of probe responses, whereas retrospective probing prioritizes the quality of the responses to the survey questions (Drennan, 2003; Fowler et al., 2016; Willis, 2005; Willis & Artino, 2013). Although standard textbooks implore the strengths and weaknesses of different probe placements (i.e., Collins, 2015, p. 120; Willis, 2005) and it is promoted to document probe placement in research reports (Boeije & Willis, 2013), theoretical discussions of the cognitive processes underlying the assumed effects of placement are lacking, as is empirical research on the effects of placement on probe response burden and response quality.

Concerns regarding response burden and the response quality of probing data are inherent to web probing. Web probes are typically administered as *open-ended narrative questions* due to their origin in cognitive interviewing. However, unlike interviewer-administered probes, web probes require that respondents type their answers autonomously (Behr et al., 2017). Consequently, web probes suffer from shorter responses, and markedly higher levels of nonresponse or otherwise uninterpretable answers than responses obtained during cognitive interviews (Lenzner & Neuert, 2017; Meitinger & Behr, 2016). It has been suggested to employ web *probes with predefined response options* (Scanlon, 2019, 2020), using single-choice answers or a check-all-that-apply (CATA) format. These closed probes cause less response burden and produce higher response quality in terms of fewer uninterpretable answers (Neuert, Meitinger,

& Behr, 2021; Scanlon, 2020). Potentially, closed probe formats are more resistant to contextual effects through probe placement.

The aim of the present research is two-fold: First, it seeks to examine the effects of probe placement on response burden and response quality of web probes. Secondly, it examines whether the effects of probe placement are moderated by probe format. No experimental research has examined probe placement and format in conjunction.

The following section discusses how probe placement and format impact the cognitive task of responding to web probes and summarizes previous research. Following this, hypotheses on the effects of probe placement and format on response burden and quality of web probes are derived, and a web experiment is reported that analyses these effects using three survey questions and probes on quality of life.

2   Background: The Technique of Probing

The technique of probing was first described and promoted by Schuman (1966) in the context of interviewer-administered surveys, in which a random subsample of respondents was asked open-ended questions about a preceding closed survey question to assess how respondents understood the survey question. The technique soon became an integral component in cognitive interviewing when pretesting draft survey questions (Beatty & Willis, 2007; Converse & Presser, 1986; Smith, 1989) as a supplement and an alternative to the think-aloud method (Fox et al., 2011; Priede & Farrall, 2011; Russo et al., 1989).

Considering that cognitive pretesting focusses on and analyses the cognitive tasks that survey questions impose on respondents, the cognitive tasks that probes impose on them receive surprisingly little attention. Probing poses an introspection-based metacognitive task (Overgaard & Sandberg, 2012), meaning respondents must self-observe and self-report their thought processes (Collins, 2003; Wilson et al., 1996). More precisely, as probes are asked *after* respondents have answered the survey question, they must retrieve information on the thought processes they had during survey response from short-term memory, referred to as retrospection (Bröder, 2019; Massen & Bredenkamp, 2005). Finally, they must translate their internal response into a verbalized or written answer (Behr et al., 2020). Answering probes is by nature a complex and burdensome task. Therefore, it is no surprise that the way probes are presented impacts the burden they place on respondents and the quality of the data collected via probing (Behr et al., 2012b).

2.1 Probe Placement: Concurrent Versus Retrospective Probing

Concurrent probing describes when a probe is presented to a respondent directly after having answered a survey question. In web probing, concurrent probes are presented on the survey page following the survey question under examination (Behr et al., 2020, p. 527f). Concurrent probing is thought not to over-burden respondents with the simultaneous tasks of answering a survey question and carrying out introspection—as is done in the think-aloud technique (Ericsson & Simon, 1980, 1993; Gerber & Wellens, 1997) and potentially when web probes are embedded on the same page as the survey question (i.e., Couper, 2013; Luebker, 2021; Neuert & Lenzner, 2023)—while ensuring that respondents' thought processes are still available in short-term memory. However, concurrent placement is not recommended by practitioners in all instances. Commonly named caveats of concurrent probing include interrupting the flow of the survey questions, particularly when multiple items or questions pertain to an overarching topic (Collins, 2015, p. 120), as this may impact how respondents process and answer subsequent survey questions (i.e., Couper, 2013; Hadler, 2023). The alternative is to place probes retrospectively, at the end of a section on an overarching topic, or the end of a survey. This, however, means that the related survey question and probe are presented at different points in the survey, interfering with the conversational logic of probing, and adversely effecting retrospection, for instance, regarding information accessibility (Drennan, 2003; Willis, 2005). Retrospective placement potentially impacts probes in two ways: it increases the perceived response burden caused by the probe and the likelihood of memory errors.

Response burden is elevated because retrospective placement asks respondents to recapitulate a foregoing survey question after having already moved on to other questions and topics. This approach contradicts the conversational *maxim of relation* (Grice, 1975), which expects that each new (survey or probing) question pertains to the previous question, thereby building and increasing *common ground* (Clark & Haviland, 1977; Schober, 1999). Response burden is often measured in terms of its negative effects on data quality, such as survey break-off (i.e., Peytchev, 2009) or item nonresponse (Holland & Christian, 2009; Miller & Lambert, 2014; Zuell, Menold, & Körber, 2015). However, direct and indirect measurements of *perceived response burden* exist (Yan & Williams, 2022). Indirect measures include signs of increased cognitive effort and reduced motivation (Yan et al., 2020). Response times are a typical measure of cognitive effort (Yan & Tourangeau, 2008). Applied to probe placement, the additional burden of retrospective probing may be visible in higher response latency, that is the time spent reading the probe and trying to re-

cap the survey question. One sign of reduced motivation to provide a high-quality response is if respondents invest less time typing the probe response when probes are asked retrospectively. Another sign is if respondents try to leave a probe unanswered altogether, thereby activating a motivational prompt (Al Baghal & Lynn, 2015; Chaudhary & Israel, 2016; Holland & Christian, 2009; Kaczmirek et al., 2017; Smyth, Dillman, Christian, & Mcbride, 2009).

With more distance between the survey question and the probe, the task of retrospection not only becomes more burdensome, but also more prone to *memory errors* (Wilson et al., 1996), meaning that participants might fail to report thoughts they had, or report ones they did not have. For one, the construal of cognitive probes depends on the information accessible to the respondent at the time of answering the probe. Due to the time lag between survey question and probe, some content may no longer be available in short-term memory, making it unreportable. This should be measurable in a lower share of interpretable probe responses, and fewer mentioned themes. For another, respondents answering probes retrospectively may be more susceptible to cues provided by the survey context to fill gaps in their memory. Respondents are *cooperative communicators* (Clark & Haviland, 1977; Grice, 1975) seeking to give relevant answers to probes, that is, answers that pertain to and support their survey response (Silber et al., 2020). When respondents cannot remember their thought processes, they tend to give answers based on theories about how one might arrive at a particular conclusion (Nisbett & Wilson, 1977). Such theories may be based on general knowledge or contextual cues, such as intermittent survey questions. For instance, if a topical block on quality of life (Felce & Perry, 1995) includes a general domain, such as life satisfaction, and several specific domains, such as relationship satisfaction or subjective health, respondents receiving a probe at the end of the survey section may falsely remember the specific domains as relevant aspects of their life satisfaction, even if they were not part of their mental construal at the time of answering the question.

Considering how central the decision of placement is when implementing probes (Willis, 2005. p. 51 f.), it is surprising how little empirical data there is on the effects of probe placement on response burden and probe response quality. The only study to date that compared concurrent and retrospective web probes is reported by Fowler and colleagues (2016; 2020). The study examined nine dichotomous items on neighbourhood walkability using four open-ended probes, implemented concurrently or retrospectively at the end of the questionnaire. Results showed a significantly higher share of relevant responses to one of four probes when placed concurrently. However, the authors describe their concurrent condition as somewhat resembling "a hybrid between concurrent and retrospective" ap-

proaches (Fowler & Willis, 2020, p. 461), as several survey questions were asked on one page, followed by a probe. Moreover, the reported study was not a randomized experiment as the conditions were fielded several weeks apart. Due to the studies' limitations in manipulation and randomization, the authors concluded that stronger effects are conceivable.

From the field of cognitive interviewing, one study found that think-aloud and concurrent probing detected a similar number of problems with survey questions, while retrospective probing uncovered markedly fewer problems (Daugherty et al., 2001). In the context of product decision-making, interviews using think-aloud generated more insights into cognitive steps and difficulties encountered during decision-making. However, retrospective probes delivered more insights into the final decision (Kuusela & Paul, 2000). In a usability study, think-aloud produced more procedural information, whereas retrospective probing produced more explanations for the final behaviour (Bowers & Snyder, 1990).

In summary, previous research on the effects of placement on probes is scarce and limited to open-ended probes. No research has empirically tested whether retrospective placement increases perceived response burden, such as an increased time needed to recapitulate the survey question, or signs of reduced motivation, for instance by taking less time to type an answer or trying to leave probes unanswered. Regarding probe response quality, studies on probe responses in web probing (Fowler & Willis, 2020) and cognitive interviewing (Bowers & Snyder, 1990; Daugherty et al., 2001; Kuusela & Paul, 2000) have delivered first evidence that retrospective placement is associated with less relevant or procedural content. Experimental designs that examine the share of interpretable answers, the amount of interpretable content, and whether intermittent survey questions contribute to memory errors by providing contextual cues are lacking.

2.2   Probe Format: Open-ended Versus "Closed" Probes

The probes in web surveys have traditionally been presented as open-ended questions due to their heritage in cognitive interviewing and its implored strength in detecting so-called silent misunderstandings and other unsuspected problems by collecting respondents' verbal reports (DeMaio & Rothgeb, 1996). Open-ended probes are often administered in the form of open-ended narrative questions with multi-line answer boxes, though single-line and adaptive text boxes are also used for probes that do not require full-sentence answers (Behr, Bandilla et al., 2014; Kunz & Meitinger, 2022). Web probes with predefined response options (Scanlon, 2019, 2020) are typically referred to as "closed" probes, though they often include an open-ended "other" field. The answer categories can be based on findings from previous cognitive interviews of even previous open-ended web probes. Currently, closed web probes are primarily used to quantify findings from cognitive interviews (i.e., Scanlon, 2019, 2020) or to carry out subgroup comparisons (i.e., Neuert, Meitinger, & Behr, 2021). The response categories in closed probes are either presented in a check-all-that-apply (CATA) or single-choice format, usually randomizing the order of the predefined responses (Neuert, Meitinger, & Behr, 2021).

Regardless of whether a probe or survey question uses an open-ended or closed format, respondents must ideally interpret the pragmatic meaning of a question, embark on the retrieval of relevant information, form an internal judgment and format their internal answer to fit the response format (Tourangeau et al., 2000). In the case of open-ended web questions, respondents perform these tasks based on the question text alone (Schuman & Presser, 1979) and autonomously type in their responses (Schmidt et al., 2020). In comparison, closed questions and probes provide response options that may contribute to the construal of a question's meaning, influence which information is retrieved, and how a judgment is formed (Schwarz et al., 1988). Because the cognitive tasks involved in answering open-ended questions are—all else equal—less defined, open-ended questions are associated with higher response burden and nonresponse. Indeed, much of the research on open-ended questions focusses on efforts to improve response quality.

Regarding perceived response burden (Yan & Williams, 2022), a study that continuously asked respondents to evaluate their survey experience found that questionnaire blocks that included open-ended narrative questions were considered more burdensome and less interesting than ones with closed questions only (Galesic, 2006). Comparing response times between open-ended and closed questions is not common due to the lack of comparability between formats, though open-ended questions are associated with longer response times. Several studies on open-ended questions have studied the effects of motivational prompts on the likelihood of giving substantive answers (Al Baghal & Lynn, 2015; Chaudhary & Israel, 2016; Holland & Christian, 2009; Kaczmirek et al., 2017; Smyth et al., 2009), as respondents are more likely to try and leave open-ended questions unanswered.

The differences between open-ended and closed web survey questions and probes regarding nonresponse and response content are well documented. The main asset of open-ended questions and probes is that respondents name a larger variety of themes and give more detailed answers (Neuert, Meitinger, & Behr, 2021; Reja et al., 2003; Zuell, 2016). However, nonresponse to open-ended

questions and probes is significantly higher, and the mean number of themes named lower as compared to closed formats (Neuert, Meitinger, & Behr, 2021; Reja et al., 2003; Schuman & Presser, 1979; Zuell et al., 2015). A study by Tourangeau et al. (2014) demonstrated that respondents' self-reports as to which types of food they had eaten were more strongly impacted by examples in the instructions when the question was asked in an open-ended than closed format. This has been interpreted as evidence that contextual information may influence open-ended questions more strongly.

In summary, while open-ended question formats provide richer and more detailed responses, they are associated with increased response burden and adverse effects on data quality, such as a higher share of nonresponse and a lower mean number of themes. Moreover, research has indicated that contextual cues impact open-ended question formats more strongly. Consequently, probes that include predefined response options may be less affected by probe placement than open-ended probes.

3 Hypotheses

The present study aims to clarify whether retrospective probe placement negatively impacts the perceived response burden and response quality of probes in web surveys and whether such effects are moderated by probe format. Based on the notion that intermittent survey questions increase response burden and memory errors, I put forward two hypotheses regarding the impact of probe placement:

Placing probes retrospectively ...

H1 ... increases the perceived response burden of probes.

H2 ... decreases probe response quality regarding probe nonresponse and probe response content.

Moreover, based on previous research on open-ended and closed survey questions and probes, I postulate that the effects of probe placement are more pronounced for open-ended probes than for probes with predefined response options:

H3 The effects of probe placement are moderated by probe format.

Regarding the first hypothesis, the perceived *response burden* is gauged with response times and the activation of motivational prompts. Response times remain a common measure of cognitive effort and response burden (Yan & Tourangeau, 2008). However, coherent response time analysis and interpretation is complex as longer response times

may indicate increased respondent motivation (Höhne, Schlosser, & Krebs, 2017) or burden (Lenzner et al., 2010). Matters are further complicated when comparing open-ended and closed question formats, as probes with predefined response options require respondents to read more text (and thus presumably spend more time reading the probe). In contrast, open-ended probes require respondents to type a response rather than simply selecting predefined response options (presumably requiring more time to respond). Due to this diminished comparability between experimental conditions regarding the total response time, the present study distinguishes between the response latency and the time spent answering, as has been done in recent studies (Meitinger, Behr, & Braun, 2019). *Response latency* is the time between the loading of the survey page and the first click or keystroke and measures the time spent reading and reflecting the probe. I expect response latency to be higher for retrospective probes (H1a) than for concurrent probes as respondents need more time to recall the survey question. Response latency should be higher for probes with predefined response options as respondents must not only read the text of the probing question but also the response options. The *time spent answering* is defined as the time between the first click/keystroke and the second to last click/keystroke (the click/keystroke before the submit button) and thus corresponds to the time spent typing in an answer to an open-ended probe or selecting the relevant response option(s). I expect the time spent answering to be longer for concurrent than retrospective probes as respondents invest more effort into their answer (H1b). Moreover, the time spent answering should be longer for open-ended probes, as typing an answer requires more clicks than selecting a response option. As a third measure of response burden, I assume that respondents are more likely to try to leave probes unanswered when they are asked retrospectively, thus activating *motivational prompts* more often (H1c). Motivational prompts state that respondents' answers are important to the purpose of the study. They have become a popular tool for increasing response quality to open-ended questions (Al Baghal & Lynn, 2015; Kaczmirek et al., 2017; Smyth et al., 2009).

Regarding Hypothesis 2 on *probe response quality*, I postulate that in retrospective probing, less content is available to respondents in their short-term memory. This should increase non-substantive probe responses (H2a) and decrease the mean number of themes (H2b) being mentioned. Furthermore, the decreased accessibility to short-term memory should make respondents more likely to use contextual information as memory cues in retrospective probing (H2c), such as cues on topically related intermittent survey questions.

Regarding the third hypothesis on the *moderating effects of probe format*, I hypothesize that the adverse effects of ret-

Table 1

Overview of hypotheses

H1: Placing probes retrospectively increases the perceived response burden of probes	
Retrospective probing …	
H1a:	… increases response latency (time before the first click/keystroke)
H1b:	… decreases the time spent answering (time between first and second-to-last click/keystroke)
H1c:	… increases the activation of motivational prompts
H2: Placing probes retrospectively decreases probe response quality	
Retrospective probing …	
H2a:	… increases the share of non-substantive probe responses (i.e., leaving a probe unanswered or providing uninterpretable content)
H2b	… decreases the mean number of themes named
H2c:	… increases the use of memory cues from intermittent survey questions
H3: The effects of probe placement are moderated by probe format	
Negative effects of retrospective probing on response burden and probe response quality are more pronounced for open-ended probes than for probes with predefined response options (interaction effect of probe placement and format) in terms of …	
H3a:	… response latency
H3b:	… the time spent answering
H3c:	… the activation of motivational prompts
H3d:	… the share of non-substantive probe responses
H3e:	… the mean number of themes
H3f:	… the use of memory cues from intermittent survey questions

rospective probing on response burden and probe response quality are more pronounced for open-ended probes than for probes with predefined response options regarding the parameters mentioned above (H3a to H3f). Thus, an interaction effect of probe placement and format is assumed. Table 1 summarizes the hypotheses.

4 Method

4.1 Experimental Design and Web Survey

A 2x2 web experiment was designed in which respondents received three questions on domains of quality of life on separate survey pages. The survey questions were either accompanied by open-ended probes or probes with predefined response options (see Fig. 1), which were presented concurrently or retrospectively. In the retrospective condition, the probes were presented after all three survey questions and several other unrelated questions. Respondents were randomly assigned to one of the four experimental conditions (see Table 2).

An online survey was conducted with a non-probability sample between November 20th and December 4th, 2020, with the panel provider Respondi AG. In total, 13,814 people were invited and 4994 respondents (36%) started the survey. Some participants were ineligible due to age or quota restrictions ($n = 301$) or did not complete the survey ($n = 307$). Of the 4386 respondents who completed the questionnaire, 2184 were part of the current experiment. The sample included quotas to depict the German online population in terms of gender (male, female)[1] and age. There were no significant differences regarding demographics or device used between experimental groups (see Table A.1 in the Appendix). Respondents received 1.00€ in incentives. Average survey completion time was 12.3 (median: 10.1) minutes.

The reported study was placed towards the beginning of the survey, after the quota-relevant questions and one other experiment (which was unrelated and assigned independently). No probes were implemented before the experiment. The three survey questions were asked directly after each other. In the conditions with concurrent probing, the probes were embedded between the survey questions on separate pages. In the conditions with retrospective probing, the survey questions were followed by an unrelated study of ten questions.[2] Then the three probes immediately followed each other (see Fig. 2).

---

[1] Respondents could also choose the non-binary category "divers"; this was however not subjected to quotas.

[2] The topics of these questions were fare evasion and knowledge questions on non-binary gender categories.

Open-ended format (P1)

The previous question was:
"How satisfied are you at present, all in all, with your life?"

On a scale from 1 to 11, with 1 indicating "not at all satisfied" and 11 indicating "totally satisfied", you answered:
„**11 totally satisfied**"

Which aspects of your life did you consider when answering the question?

Closed format (P1)

The previous question was:
"How satisfied are you at present, all in all, with your life?"

On a scale from 1 to 11, with 1 indicating "not at all satisfied" and 11 indicating "totally satisfied", you answered:
„**11 totally satisfied**"

Which aspects of your life did you consider when answering the question?

☐ my family life

☐ my relationship

☐ my work life

☐ my free time

☐ my health

☐ other (please insert):

Fig. 1

Probe format

Table 2

Experimental conditions

	Probe format	
Probe placement	Open (Open-ended text field)	Closed (Predefined response options)
Concurrent	A: Open-ended, concurrent	C: Closed, concurrent
Retrospective	B: Open-ended, retrospective	D: Closed, retrospective

The Universal Client-Side Paradata script by Kaczmirek and Neubarth (2007) was implemented to ensure an exact measure of response times (Yan & Tourangeau, 2008) and collect questionnaire navigation data (Callegaro et al., 2015; Kunz & Hadler, 2020), such as the activation of motivational prompts. Following legal and ethical research standards (ADM, ASI, BVM, & DGOF, 2021; Kunz, Beuthner, Hadler, Roßmann, & Schaurer, 2020), respondents were informed about the collection and use of client-side paradata on the welcome page of the survey.

4.2 Survey Questions and Probes

The survey questions comprised three measures of quality of life (Felce & Perry, 1995; Theofilou, 2013; Veenhoven, 2000) consisting of one general assessment and two specific domains. The general measure was a question on life satisfaction (Q1) (Beierlein et al., 2014) using an 11-point scale ranging from 1, "not at all satisfied" to 11, "totally satisfied" and including an explicit nonresponse option "I do not want to answer". The second question asked about the domain of relationship satisfaction (Q2) employing the same response options (Schwarz, Strack, & Mai, 1991). The third was a measure of subjective health (Q3) with a five-point scale ranging from "very good" to "very poor" (De Bruin et al., 1996). There were no significant differences in re-

A: Open-ended, concurrent	B: Open-ended, retrospective	C: Closed, concurrent	D: Closed, retrospective
Q1: Life satisfaction	Q1: Life satisfaction	Q1: Life satisfaction	Q1: Life satisfaction
P1: Probe Life satisfaction (open-ended)	Q2: Relationship satisfaction	P1: Probe Life satisfaction (closed)	Q2: Relationship satisfaction
Q2: Relationship satisfaction	Q3: Subjective health	Q2: Relationship satisfaction	Q3: Subjective health
P2: Probe Relationship satisfaction (open-ended)	Unrelated questions	P2: Probe Relationship satisfaction (closed)	Unrelated questions
Q3: Subjective health	P1: Probe Life satisfaction (open-ended)	Q3: Subjective health	P1: Probe Life satisfaction (closed)
P3: Probe Subjective health (open-ended)	P2: Probe Relationship satisfaction (open-ended)	P3: Probe Subjective health (closed)	P2: Probe Relationship satisfaction (closed)
Unrelated questions	P3: Probe Subjective health (open-ended)	Unrelated questions	P3: Probe Subjective health (closed)

Fig. 2

Study design

sponse distributions or item nonresponse between experimental conditions for any survey questions.

Each survey question was accompanied by a specific probe, which repeated the question text and the respondent's answer, and asked which aspects of their life (P1), relationship (P2), or health (P3) they had considered when answering the question. Probing questions were worded identically across all conditions. The open-ended probes included an open-ended text field. The probes with predefined response options presented these in a check-all-that-apply (CATA) format with an open-ended "other" option at the bottom. The order of the predefined response options

was randomized (see Appendix A.2 for the original survey questions and probes and an English translation). Respondents who tried to leave a probe unanswered were prompted to respond using a motivational statement ("This question is very important.").

4.3 Predefined Probe Response Options and Coding of Open-ended Probe Responses

For the probe on life satisfaction (P1), the predefined response options included the two specific domains of rela-

tionship and health (Lee, McClain, Webster, & Han, 2016; Schwarz, Strack, & Mai, 1991), as well as other known correlates of life satisfaction such as job, leisure time and family life satisfaction (Theofilou, 2013). The predefined probe responses for relationship satisfaction (P2) were based on the dimensions of intimacy, passion, and commitment in line with Sternberg's (1997) triangular theory of love and augmented by relationship status based on previous research (Hadler, 2023). The predefined categories for the probe on subjective health (P3) were based on the existing codes of Lee et al. (2020), adapted to the German context and reduced to include a similar number of response options as the previous two probes.

The predefined response options were used as codes for corresponding responses in the open-ended probes. Additional themes that emerged during coding were established using an inductive approach (Willis, 2015a). Themes named by 20 or more respondents were maintained as distinct themes; all others were summarized under "other". This resulted in nine additional themes for the first and third probes, eight for the second, and the "other" category for all probes. The complete coding schemes are in Table A.3 of the Appendix.

Probe responses were coded as non-substantive when they contained only uninterpretable content. For open-ended probes, this was the case when respondents left the text field empty, inserted random characters, refusals, "don't know" answers, repeated their survey response, gave an off-topic answer, or an answer so ambiguous or vague that it could not be coded to pertain to a substantive code (i.e. "I thought of all aspects of my life") (Behr, Braun et al., 2014; Naber & Padilla, 2022). Probe responses in CATA format were marked as non-substantive when respondents did not select any of the predefined response options, or only selected the open-ended "other" category and inserted an uninterpretable response.

All open-ended probe responses were independently coded as substantive or non-substantive by the author and a second researcher, with Cohen's Kappa of 0.948 (P1), 0.856 (P2), and 0.921 (P3). The author and a student assistant independently coded the substantive responses. For the predefined categories, an intercoder reliability of 0.980–1.000 (P1), 0.832–0.987 (P2) and 0.867–1.000 (P3) was reached. For the additional themes that emerged, Cohen's Kappa ranged from 0.896–0.992 (P1), 0.841–0.930 (P2) and 0.778–0.969 (P3). Differences in codes were discussed and final codes were assigned together. The response distributions of all predefined and additional themes across experimental conditions can be found in Tables A.4 and A.5 in the Appendix.

4.4 Data Preparation and Analysis

All analyses employed probe placement (1 = concurrent; 2 = retrospective) and format (1 = open-ended; 2 = predefined response options) as main predictors, with probe placement used to test the first and second hypotheses. All two-way models included an interaction of probe placement and format to test the third hypothesis. Gender, age, education, and device type were included as covariates. The analyses of motivational prompts and share of non-substantive responses were carried out based on all probe responses. All other analyses were carried out based on substantive probe responses only.

Dichotomous dependent variables were examined when possible using binary logistic regression with the main predictors and covariates described above. This was the case for the share of non-substantive response options (1 = substantive probe response; 0 = non-substantive probe response) and the prevalence of the specific domains "relationship" and "health" from Q2 and Q3 in answer to the probe on the general domain of life satisfaction (P1; 1 = content named; 0 = content not named). Unfortunately, the low prevalence of activated motivational prompts did not permit carrying out regression analysis for this parameter. Therefore, Pearson's chi-square tests of independence are reported.

Metric dependent variables were response times and the number of themes. They were examined using multivariate analyses of covariance (MANCOVA) across the three probes with the main predictors and covariates as described above. Response time data is positively skewed and subject to outliers; therefore, outliers must be defined, handled (i.e., omitted or replaced with other values) and transformed prior to analysis (Kunz & Hadler, 2020). Various response time outlier definitions exist (Matjašič, Vehovar, & Lozar Manfreda, 2018). In the present study, outliers were excluded using Tukey's method ($Q_{.25} - 1.5$ IQR / $Q_{.75} + 1.5$ IQR) (Tukey, 1977), as researchers have increasingly recommended basing outlier definitions on the median, quartiles and interquartile ranges (IQR) rather than on the mean, which is more strongly impacted by outliers (i.e., Höhne & Schlosser, 2018). Tukey's method led to between 6% and 9% of response times being identified as outliers. Outliers were set to missing and valid response time data was log-transformed. MANCOVAs were carried out with the valid and log-transformed response time data of the substantive probe responses for response latency and time spent answering. The robustness of response time analyses was tested by applying an alternative outlier definition (Revilla & Couper, 2018), which only excluded response times beyond the upper and lower one percentile and log-transformed the remaining data (Yan & Tourangeau, 2008).

Table 3

Probe response times, MANCOVAs

Main predictors	Response latency				Time spent answering			
	Wilk's λ	F	p	η^2	Wilk's λ	F	p	η^2
Placement*format	0.98	9.62[a]	<0.001	0.02	0.99	2.38[b]	0.068	–
Probe placement	0.94	27.91[a]	<0.001	0.06	1.00	1.08[b]	0.356	–
Probe format	0.67	215.66[a]	<0.001	0.33	0.59	306.77[b]	<0.001	0.41
Between-subjects effects								
Placement*format								
P1	–	1.29[c]	0.255	–	–	–	–	–
P2	–	11.18[c]	0.001	0.01	–	–	–	–
P3	–	10.65[c]	0.001	0.01	–	–	–	–
Probe placement								
P1	–	52.48[c]	<0.001	0.04	–	–	–	–
P2	–	50.17[c]	<0.001	0.04	–	–	–	–
P3	–	62.25[c]	<0.001	0.05	–	–	–	–
Probe format								
P1	–	33.23[c]	<0.001	0.02	–	760.39[d]	<0.001	0.37
P2	–	294.01[c]	<0.001	0.18	–	394.00[d]	<0.001	0.23
P3	–	540.55[c]	<0.001	0.29	–	488.88[d]	<0.001	0.27
N	1308				1332			

a df = 3 and 1298, *b* df = 3 and 1222, *c* df = 1 amd 1300, *d* df = 1 and 1324

All analyses revealed the same overall effects; differences in between-subject effects are discussed where applicable.

All analyses were carried out using IBM SPSS Statistics Version 24.0.

5   Results

5.1   Response Burden

The first hypothesis predicted that the response burden would be higher when retrospective probing is used, resulting in increased response latency (H1a), decreased time spent answering (H1b), and increased activation of motivational prompts (H1c). The third hypothesis predicted that these effects would be more pronounced for open-ended probes than for probes with predefined response options (H3a to H3c).

Response times. After excluding response time outliers, 1308 cases remained for the MANCOVA of *response latency* (concurrent: *n* = 695; retrospective: *n* = 613). There was a significant but small interaction of probe placement and format, supporting H3a, a significant, medium main effect of probe placement, supporting H1a, and a strong and

significant effect of probe format (see Table 3). Gender, age, and device were significant covariates, whereas education was not.[3] Fig. 3 depicts the mean response latencies and standard deviations for the three probes after outlier exclusion. Response latency was higher when probes were placed retrospectively and when they included predefined response options, so the main effects remain interpretable while an overall interaction effect exists. The between-subjects effects confirm significant but minimal interaction effects for the probes on relationship satisfaction (P2) and subjective health (P3) but not for life satisfaction (P1). A MANCOVA based on response times that only excluded the top and bottom percentile showed the same overall effects; however, the between-subjects effects showed the opposite pattern of effects, with a significant interaction for the probe on life satisfaction (P1), but not for the other two probes. The effect sizes of the interactions remained negligible across all analyses, so that the interaction effect cannot be interpreted substantively.

---

[3] MANCOVA results with respect to the covariates (1) gender: Wilks-Lambda = 0.99; $F_{(3,1298)}$ = 5.40; p = 0.001; η^2 = 0.01; (2) age: Wilks-Lambda = 0.87; $F_{(3,1289)}$ = 67.29; p < 0.001; η^2 = 0.13; (3) education: Wilks-Lambda = 1.00; $F_{(3,1298)}$ = 0.88; p = 0.453 n.s.; (4) device used: Wilks-Lambda = 0.96; $F_{(3,1298)}$ = 18.77; p < 0.001; η^2 = 0.04.
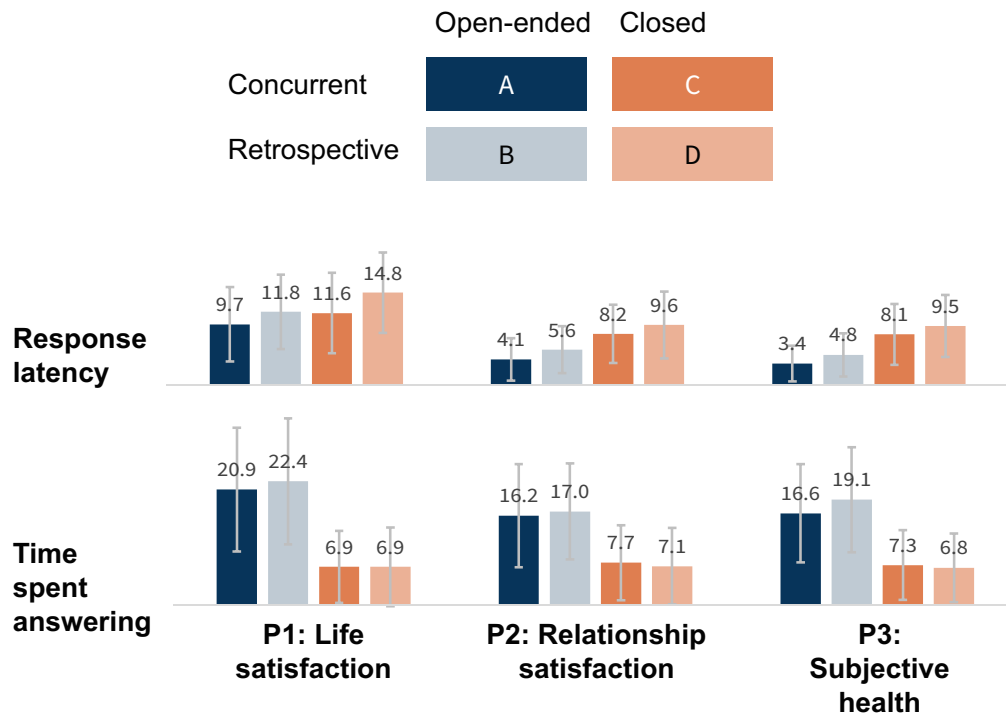
Fig. 3

Probe response times (mean and standard deviation)

Regarding the *time spent answering*, 1332 cases were included in the MANCOVA (concurrent: $n = 676$; retrospective: $n = 656$). There was no significant main effect of probe placement and the interaction effect of probe placement and format failed to reach significance (Table 3), lending no support to H1b or H3b. Again, the probe format exerted a strong and significant influence. Age was the only significant covariate.[4] A MANCOVA based on the alternative response time outlier exclusion confirmed the overall and between-subjects effects.

The lower row of Fig. 3 shows that respondents took markedly longer to type their responses to open-ended probes than to select the appropriate response option(s) in the check-all-that-apply format. Based on the descriptive data, respondents took slightly longer to answer open-ended probes in the retrospective condition across all three probes (contrary to expectations); however, this tendency did not reach significance in any of the analyses performed.

Motivational prompts. In total, only 69 (3%) respondents tried to leave one or several probes unanswered and received a motivational prompt, so binary logistic regression and testing for an interaction of probe placement and format was not possible. However, the prevalence across experimental groups showed that the likelihood of trying to leave a probe unanswered did not differ by probe placement (concurrent: 3%; retrospective: 3%; $\chi^2(1) = 0.010$; $p = 0.921$), whereas open-ended probes were significantly more likely to be associated with activating prompts than probes with predefined response options (open-ended: 5%; closed: 1%; $\chi^2(1) = 30.733$; $p < 0.001$).

5.2 Probe Response Quality

The second hypothesis predicted that probe response quality would be lower for retrospectively placed probes, resulting in a higher share of non-substantive probe responses (H2a), a lower mean number of themes named (H2b), and an increased reliance on memory cues from the intermittent survey questions on relationship satisfaction and subjective health while responding to the probe on life satisfaction (P1) (H2c). The third hypothesis predicted that these effects would be more pronounced for open-ended probes than for probes with predefined response options (H3d to H3f).

---

[4] MANCOVA results with respect to the covariates (1) gender: Wilks-Lambda = 1.00; $F_{(3,1322)} = 2.09$; $p = 0.010$; n.s.; (2) age: Wilks-Lambda = 0.96; $F_{(3,1322)} = 16.64$; $p < 0.001$; $\eta^2 = 0.04$; (3) education: Wilks-Lambda = 1.00; $F_{(3,1322)} = 1.84$; $p = 0.138$ n.s.; (4) device used: Wilks-Lambda = 1.00; $F_{(3,1322)} = 1.56$; $p = 0.197$; n.s.

Table 4

Non-substantive probe responses, binary logistic regressions

	P1: Life satisfaction		P2: Relationship satisfaction		P3: Subjective health	
	%	*n*	*%*	*n*	*%*	*n*
A: Open-ended, concurrent	20	106	31	168	28	153
B: Open-ended, retrospective	25	135	35	188	31	168
C: Closed, concurrent	2	9	2	10	2	8
D: Closed, retrospective	4	19	3	16	2	13
Binary logistic regression	OR		OR		OR	
Placement*format	0.61		0.70		0.69	
Probe placement	1.69*		1.36		1.33	
Probe format	0.08***		0.05***		0.04***	
Model χ^2 (7)	258.35***		442.83***		428.35***	
Correct classification (%)	87.7		82.5		84.7	
Nagelkerke R^2	0.213		0.304		0.308	
N (Basis: all probe responses)	2181		2181		2181	

*$p < 0.05$; **$p < 0.01$; ***$p < 0.001$

Table 5

Mean number of themes, descriptive results

	P1: Life satisfaction		P2: Relationship satisfaction		P3: Subjective health	
Mean number of themes	Mean	SD	Mean	SD	Mean	SD
A: Open, concurrent	2.50	1.42	1.60	1.06	1.43	0.77
B: Open, retrospective	2.52	1.39	1.52	0.97	1.46	0.84
C: Closed, concurrent	2.73	1.45	2.46	1.71	2.45	1.51
D: Closed, retrospective	2.63	1.50	2.28	1.65	2.30	1.45
N (Basis: substantive probe responses)	1915		1802		1842	

Non-substantive probe responses. Table 4 shows the share of non-substantive responses by probe placement and format for all three probes. The share of non-substantive responses was much higher for open-ended probes (between 20% and 35%) than for probes in the check-all-that-apply format (between 2% and 4%). Across both probe formats, the share of non-substantive responses was slightly higher in the retrospective conditions based on the descriptive data.

A binary logistic regression was performed for each probe. All models were statistically significant, explained between 21% and 31% (Nagelkerke R^2) of the variance in non-substantive responding, and correctly classified over 80% of cases. Retrospective probes were associated with an increase in the likelihood of providing a non-substantive response for the probe on life satisfaction (P1) only (OR = 1.69, 95% CI [1.10, 2.59]), partially confirming H2a. There was no interaction effect of probe placement and format for any examined probes, lending no support for H3d. Open-ended probes were associated with a substantial increase in the likelihood of providing a non-substantive response for all probes. Women were more likely to offer substantive content than men for all probes; age and education were significant covariates for the probes on relationship satisfaction (P2) and subjective health (P3), and the device was a significant covariate for the probe on relationship satisfaction (P2) only.

Mean number of themes. Table 5 shows the mean number of themes for each probe and condition. For the probe on life satisfaction (P1), the mean number of themes was similar across all four conditions (between 2.50 and 2.73), while for the other two probes, open-ended probes produced a markedly lower mean number of themes (between 1.43 and 1.60) than probes with predefined response options (between 2.28 and 2.46).

Table 6

Mean number of themes, MANCOVA

Main predictors	Mean number of themes			
	Wilk's λ	$F_{(3,1600)}$	p	η^2
Placement*format	1.00	0.89[a]	0.443	–
Probe placement	1.00	1.00[a]	0.391	–
Probe format	0.87	79.46[a]	< 0.001	0.13
Between-subjects effects for significant predictors				
Probe format	–			
P1	–	0.40[b]	0.529	–
P2	–	90.24[b]	< 0.001	0.05
P3	–	160.97[b]	< 0.001	0.09
N	1610			

a df = 3 and 1600, *b* df = 1 and 1602

Table 7

Themes "relationship" and "health", binary logistic regressions

	Relationship		Health	
	%	n	%	n
A: Open-ended, concurrent	16	70	41	177
B: Open-ended, retrospective	26	106	38	156
C: Closed, concurrent	49	263	64	342
D: Closed, retrospective	48	253	58	310
Binary logistic regression	OR		OR	
Placement*format	0.474***		0.923	
Probe placement	0.732**		1.189	
Probe format	0.264***		0.419***	
Model χ^2 (7)	232.63***		115.23***	
Correct classification (%)	67.6		61.2	
Nagelkerke R^2	0.157		0.078	
N (Basis: substantive probe responses)	1913		1913	

*p < 0.05; **p < 0.01; ***p < 0.001

After excluding non-substantive responses, 1610 cases remained for the MANCOVA of the number of themes (concurrent: *n* = 817; retrospective: *n* = 793). There were no significant effects of probe placement, nor an interaction of probe placement and format (Table 6), lending no support to H2b or H3e. Probe format exerted a strong and significant main effect, with respondents selecting more themes in the conditions with predefined response options. Gender, age, and education were significant covariates, whereas the device was not.[5] The test of between-subjects effects confirmed the descriptive results that probe format was a significant predictor of the number of themes for the probes on relationship satisfaction (P2) and subjective health (P3) of medium effect size but not for the general domain of life satisfaction (P1).

---

[5] MANCOVA results with respect to the covariates (1) gender: Wilks-Lambda = 0.99; $F_{(3,1600)}$ = 3.10; p = 0.027; η^2 = 0.01; (2) age: Wilks-Lambda = 0.97; $F_{(3,1600)}$ = 15.82; p < 0.001; η^2 = 0.03; (3) education: Wilks-Lambda = 1.00; $F_{(3,1298)}$ = 9.22; p < 0.001; η^2 = 0.02; (4) device used: Wilks-Lambda = 1.00; $F_{(3,1600)}$ = 0.40; p = 0.750; n. s.

Table 8

Summary of results

	Interaction of probe placement and format	Main effect of probe placement	Main effect of probe format
Response burden			
Response latency	Yes, but minimal effect size	Yes	Yes
Time spent answering	No	No	Yes
Motivational prompts	n. a.	No	Yes
Response quality			
Non-substantive probe responses	No	Partially (P1)	Yes
Mean number of themes	No	No	Yes
Reliance on memory cues (P1)	Partially (theme "relationship")	Partially (theme "relationship")	Yes

Reliance on memory cues from intermittent survey questions. Whereas the question on life satisfaction depicts a general measure of quality of life, the subsequent questions on relationship satisfaction and subjective health focus on specific domains that may or may not be relevant to a person's overall life satisfaction. Respondents in the concurrent condition received the probe asking them to name relevant aspects of their life satisfaction (P1) *before* answering the questions on specific domains. In contrast, respondents in the retrospective condition received this probe *after* the survey questions on the specific domains. Based on the notion that respondents have less access to their short-term memory in retrospective probing and rely more heavily on contextual information as memory cues, Hypothesis 2c postulated that the themes "relationship" and "health" were more likely to be named in retrospective conditions, and Hypothesis 3f that this effect would be stronger for the open-ended probe. Table 7 shows the prevalence of the two themes by experimental condition and binary logistic regressions for each theme. Both models were statistically significant, explained between 8% and 16% (Nagelkerke R^2) of the variance in mentioning the respective theme, and correctly classified over 60% of cases.

Regarding the likelihood of mentioning the theme "relationship" as a relevant aspect of one's life satisfaction, there was a significant interaction of probe placement and format (OR = 0.47, 95% CI [0.31, 0.72]), as well as significant main effects of predictors. In the open-ended condition, only 16% of respondents named the theme "relationship" when the probe was asked concurrently, whereas 26% did this in the retrospective condition when they were intermittently presented the survey question on relationship satisfaction. Mentioning the theme "relationship" occurred significantly more often in the conditions with predefined response options; however, within the closed conditions, there was no significant difference based on probe placement (concurrent: 49%; retrospective: 48%).

In contrast, for the model of the theme "health", there was no significant interaction of probe placement and format, nor did the main effect of probe placement reach significance (OR = 1.19, p = 0.068 n. s.). Probe format was associated with an increased likelihood of mentioning the theme. Thus, Hypotheses 2c and 3f can be confirmed for the theme "relationship" but not for "health".

6   Discussion and Conclusion

The present study was designed to determine the effects of concurrent and retrospective probe placement on response burden and response quality of web probing data, and whether these effects are moderated by probe format. To this purpose, a 2x2 web experiment was designed that randomly assigned respondents to conditions with concurrent or retrospective probes that employed an open-ended response format or included predefined response options.

The hypotheses that retrospective probing increases perceived response burden (H1) and that this effect is moderated by probe format (H3) were confirmed for response latency only. Placing probes retrospectively increased the time between loading the survey page containing the probe and the first click or keystroke. This indicates that respondents need longer to recapitulate survey questions when probes do not directly follow them but are asked later in the questionnaire. The interaction of probe placement and format regarding response latency was significant, but so small in size that it forbids substantive interpretation. Contrary to the first hypothesis, probe placement did not affect the time respondents invested in answering the probes. There was no empirical support for the notion that retrospective probe placement increases the likelihood of respondents trying to leave probes unanswered, activating motivational prompts.

The second hypothesis that retrospective probing decreases probe response quality and the third hypothesis that this effect is moderated by probe format were partially con-

firmed. The share of non-substantive responses was significantly increased by retrospective placement for the probe on life satisfaction; however, this effect was not moderated by probe format. Importantly, retrospective probe placement and format impacted probe response content in one case. Respondents who received the probe on life satisfaction in an open-ended format and retrospectively were significantly more likely to name their relationship as a relevant aspect of their life satisfaction. This indicates that respondents relied on a memory cue from the intermittent survey question on relationship satisfaction when answering the probe on life satisfaction. There was no effect of probe placement on the likelihood of mentioning this theme when respondents received the probe with predefined response options. Moreover, there was no effect of probe placement on the likelihood of mentioning subjective health, the topic of the other intermittent survey question.

In summary, probe placement impacted three indicators of response burden and quality, those being response latency, the share of substantive answers (for the probe on life satisfaction), and the reliance on memory cues (for the topic "relationship"). The effect of probe placement on the reliance on memory cues was moderated by probe format. Consistent with previous research on open-ended probes and other open-ended questions in web surveys, the response burden was higher for open-ended probes than for those with predefined response options (Galesic, 2006) and response quality was decreased in terms of nonresponse and number of themes named (Neuert, Meitinger, & Behr, 2021; Reja et al., 2003). The results of the study are summarized in Table 8.

There are at least four potential limitations concerning the generalizability of the results of this study. First, the effect of probe placement depends on its operationalization, that is the distance between the retrospective probes and the survey questions they pertain to. The present study inserted several unrelated questions between the survey questions and retrospective probes. This was done to avoid overly strong effects of the specific domains relationship and health on the probe on life satisfaction. Probes were not placed at the very end of the questionnaire to avoid overly strong effects of probe position. While this is a reasonable compromise for the research purpose, it should be noted that researchers employing other designs may encounter slightly different results. For instance, in the present study, there was a tendency towards a higher share of non-substantive responses in the retrospective conditions for all probes. However, this was only significant for the probe on life satisfaction. Possibly, web probing designs that implement probes directly following the thematic block of questions (in this case, the three measures of quality of life) would experience no increase in non-substantive responding at all. Similarly, web probing designs that place

retrospective probes at the end of a lengthy questionnaire might find more significant increases in non-substantive responses for all probes. Moreover, how strongly intermittent survey questions are used as memory cues may depend on how close retrospective probes are to the topically related, intermittent questions.

Second, each survey question was examined using one probe. The effects of probe placement and format may differ when a survey question is followed by several probes, a design known to cause a high respondent burden (Meitinger et al., 2022). Third, the present study used a narrow thematic range (measures of quality of life) and one probe type (specific probes) only. Finally, the order of the three tested survey questions and probes was not randomized. The order had to be fixated to examine the effects of the specific domains relationship and health on the first-shown question on life satisfaction. At the same time, this research design decision means that probe placement and position were not perfectly separated (Behr et al., 2012a; Neuert & Lenzner, 2021).

Despite these limitations, the study has several practical implications. Researchers should employ retrospective probing sparingly. Respondents need longer to recapitulate a survey question when a probe is asked later in the survey than when it directly follows the survey question. This increased response burden may result in a higher share of non-substantive probe responses and a higher proportion of memory errors, with respondents relying on contextual cues to answer open-ended probes. Employing probes with predefined response options rather than open-ended probes diminished the effect of intermittent survey questions on probe response content. However, the adverse effects on response latency and non-substantive probe responses occurred in both probe formats. At the same time, researchers should view the results of this study in conjunction with other research on web probing. For instance, concurrent probes have been shown to impact response times and response behaviour for subsequent, related survey questions in other studies (Hadler, 2023), and several probes about one survey question or topic may impact each other (Hadler, 2021; Meitinger et al., 2018).

Thus, while the present study enhances our understanding of the impact of probe placement and format on the perceived response burden and response quality of web probes, decisions on optimal web probing design will continue to depend on researchers' analytic focus. The present study hopefully contributes valuable insights to the growing empirical data on optimal probe implementation in web surveys.

References

ADM, ASI, BVM, & DGOF (2021). Richtlinie für Online-Befragungen. https://www.adm-ev.de/wp-content/uploads/2021/03/RL-Online-2021-neu.pdf

Al Baghal, T., & Lynn, P. (2015). Using motivational statements in web-instrument design to reduce item-missing rates in a mixed-mode context. *Public Opinion Quarterly*, *79*(2), 568–579. https://doi.org/10.1093/poq/nfv023.

Beatty, P. C., & Willis, G. B. (2007). Research synthesis: the practice of cognitive interviewing. *Public Opinion Quarterly*, *71*(2), 287–311. https://doi.org/10.1093/poq/nfm006.

Behr, D., Braun, M., Kaczmirek, L., & Bandilla, W. (2012a). Testing the Validity of Gender Ideology Items by Implementing Probing Questions in Web Surveys. *Field Methods*, *25*(2), 124–41. https://doi.org/10.1177/1525822X12462525.

Behr, D., Kaczmirek, L., Bandilla, W., & Braun, M. (2012b). Asking probing questions in web surveys: which factors have an impact on the quality of responses? *Social Science Computer Review*, *30*(4), 487–498. https://doi.org/10.1177/0894439311143530 5.

Behr, D., Bandilla, W., Kaczmirek, L., & Braun, M. (2014). Cognitive probes in web surveys: on the effect of different text box size and probing exposure on response quality. *Social Science Computer Review*, *32*(4), 524–533. https://doi.org/10.1177/0894439313485203.

Behr, D., Braun, M., Kaczmirek, L., & Bandilla, W. (2014). Item comparability in cross-national surveys: results from asking probing questions in cross-national web surveys about attitudes towards civil disobedience. *Quality & Quantity*, *48*(1), 127–148. https://doi.org/10.1007/s11135-012-9754-8.

Behr, D., Meitinger, K., Braun, M., & Kaczmirek, L. (2017). *Web probing—implementing probing techniques from cognitive interviewing in web surveys with the goal to assess the validity of survey questions. GESIS Survey Guidelines*. Mannheim: GESIS—Leibniz Institute for the Social Sciences. https://doi.org/10.15465/gesis-sg_en_023.

Behr, D., Meitinger, K., Braun, M., & Kaczmirek, L. (2020). Cross-national web probing: an overview of its methodology and its use in cross-national studies. In P. C. Beatty, D. Collins, L. Kaye, J.-L. Padilla, G. B. Willis & A. Wilmot (Eds.), *Advances in questionnaire design, development, evaluation and testing* (pp. 521–543). Hoboken: John Wiley & Sons.

Beierlein, C., Kovaleva, A., László, Z., Kemper, C. J., & Rammstedt, B. (2014). *Kurzskala zur Erfassung der Allgemeinen Lebenszufriedenheit (L-1)*. https://doi.org/10.6102/zis229. Zusammenstellung sozialwissenschaftlicher Items und Skalen (ZIS)

Boeije, H., & Willis, G. B. (2013). The cognitive interviewing reporting framework (CIRF). *Methodology*, *9*(3), 87–95. https://doi.org/10.1027/1614-2241/a000075.

Bowers, V. A., & Snyder, H. L. (1990). Concurrent versus retrospective verbal protocol for comparing window usability. *Proceedings of the Human Factors Society Annual Meeting*, *34*(17), 1270–1274. https://doi.org/10.1177/154193129003401720.

Bröder, A. (2019). Methods for studying human thought. In R. J. Sternberg & J. Funke (Eds.), *The psychology of human thought: an introduction* (pp. 27–53). Heidelberg: Heidelberg University Publishing.

de Bruin, A., Picavet, H. S. J., & Nossikov, A. (1996). Health interview surveys: towards international harmonization of methods and instruments. https://apps.who.int/iris/handle/10665/107328 WHO regional publications, European series: vol. 58.

Callegaro, M., Lozar Manfreda, K., & Vehovar, V. (2015). *Web survey methodology*. Los Angeles: SAGE.

Chaudhary, A., & Israel, G. D. (2016). Assessing the influence of importance prompt and box size on response to open-ended questions in mixed mode surveys: evidence on response rate and response quality. *Journal of Rural Social Sciences*, *31*(3), 140–159. Retrieved from https://egrove.olemiss.edu/jrss/vol31/iss3/7.

Clark, H. H., & Haviland, S. E. (1977). Comprehension and the given-new contract. In R. O. Freedle (Ed.), *Discourse production and comprehension* (pp. 1–40). Norwood: Ablex Publishing Corporation.

Collins, D. (2003). Pretesting survey instruments: an overview of cognitive methods. *Quality of Life Research*, *12*(3), 229–238. https://doi.org/10.1023/A:1023254226592.

Collins, D. (Ed.). (2015). *Cognitive interviewing practice*. London: SAGE.

Converse, J. M., & Presser, S. (1986). *Survey questions: handcrafting the standardized questionnaire*. Iowa: SAGE.

Couper, M. P. (2013). Research note: Reducing the threat of sensitive questions in online surveys? *Survey Methods: Insights from the Field*. https://doi.org/10.13094/SMIF-2013-00008.

Daugherty, S., Harris-Kojetin, L., Squire, C., & Jaël, E. (2001). Maximizing the quality of cognitive interviewing data: an exploration of three approaches and their informational contributions. https://www.researchgate.net/profile/Lauren-Harris-Kojetin/publication/266866573_MAXIMIZING_T

HE_QUALITY_OF_COGNITIVE_INTERVIEWIN
G_DATA_AN_EXPLORATION_OF_THREE_APP
ROACHES_AND_THEIR_INFORMATIONAL_C
ONTRIBUTIONS/links/54d108b90cf25ba0f0409c
5a/MAXIMIZING-THE-QUALITY-OF-COGNITIV
E-INTERVIEWING-DATA-AN-EXPLORATION-O
F-THREE-APPROACHES-AND-THEIR-INFOR-
MATION-CONTRIBUTIONS.pdf Proceedings
of the annual meeting of the American Statistical
Association.

DeMaio, T., & Rothgeb, J.M. (1996). Cognitive inter-
viewing techniques: In the lab and in the field.
In I.N. Schwarz & S. Sudman (Eds.), *Answering
questions: Methodology for determining cognitive
and communicative processes in survey research*
(pp. 177–195). San Fransisco: Jossey-Bass.

Drennan, J. (2003). Cognitive interviewing: verbal data in
the design and pretesting of questionnaires. *Journal
of Advanced Nursing*, *42*(1), 57–63. https://doi.org/
10.1046/j.1365-2648.2003.02579.x.

Edgar, J., Murphy, J., & Keating, M.D. (2016). Com-
paring traditional and crowdsourcing methods for
pretesting survey questions. *SAGE Open*, *6*(4), 1–14.
https://doi.org/10.1177/2158244016671770.

Ericsson, K.A., & Simon, H.A. (1980). Verbal reports as
data. *Psychological Review*, *87*(3), 215–251. https://
doi.org/10.1037/0033-295X.87.3.215.

Ericsson, K.A., & Simon, H.A. (1993). *Protocol analysis:
verbal reports as data* (2nd edn.). Cambridge: MIT
Press.

Felce, D., & Perry, J. (1995). Quality of life: its defini-
tion and measurement. *Research in Developmental
Disabilities*, *16*(1), 51–74. https://doi.org/10.1016/
0891-4222(94)00028-8.

Foddy, W. (1998). An empirical evaluation of in-depth
probes used to pretest survey questions. *Sociolog-
ical Methods & Research*, *27*(1), 103–133. https://
doi.org/10.1177/0049124198027001003.

Fowler, S.L., & Willis, G.B. (2020). The practice of
cognitive interviewing through web probing. In
P.C. Beatty, D. Collins, L. Kaye, J.-L. Padilla,
G.B. Willis & A. Wilmot (Eds.), *Advances in ques-
tionnaire design, development, evaluation and test-
ing* (pp. 451–469). Hoboken: John Wiley & Sons.

Fowler, S.L., Willis, G.B., Moser, R.P., Townsend, R.L.M.,
Maitland, A., Sun, H., & Berrigan, D. (2016). *Web
probing for question evaluation: the effects of probe
placement*. AAPOR. The American Association for
Public Opinion Research (AAPOR) 71st Annual
Conference.

Fox, M.C., Ericsson, K.A., & Best, R. (2011). Do pro-
cedures for verbal reporting of thinking have to
be reactive? A meta-analysis and recommendations

for best reporting methods. *Psychological Bulletin*,
137(2), 316–344. https://doi.org/10.1037/a0021663.

Galesic, M. (2006). Dropouts on the web: effects of inter-
est and burden experienced during an online sur-
vey. *Journal of Official Statistics*, *22*(2), 313–328.
Retrieved from https://www.scb.se/contentassets/
f6bcee6f397c4fd68db6452fc9643e68/dropouts-on-
the-web-effects-of-interest-and-burden-experienced-
during-an-online-survey.pdf.

Gerber, E.R., & Wellens, T.R. (1997). Perspectives on
pretesting: "cognition" in the cognitive interview?
Bulletin de Méthodologie Sociologique, *55*, 18–39.
Retrieved from https://journals.sagepub.com/doi/
pdf/10.1177/075910639705500104.

Grice, H.P. (1975). Logic and conversation. In P. Cole &
J.L. Morgen (Eds.), *Syntax and semantics: volume 3:
speech acts* (5th edn., pp. 41–58). New York: Aca-
demic Press.

Hadler, P. (2021). Question order effects in cross-cul-
tural web probing—pretesting behavior and at-
titude questions. *Social Science Computer Re-
view*, *39*(6), 1292–1312. https://doi.org/10.1177/
0894439321992779.

Hadler, P. (2023). The effects of open-ended probes on
closed survey questions in web surveys. *Sociolog-
ical Methods & Research*. https://doi.org/10.1177/
00491241231176846.

Höhne, J.K., & Schlosser, S. (2018). Investigating the ad-
equacy of response time outlier definitions in com-
puter-based web surveys using paradata surveyfocus.
Social Science Computer Review, *36*(3), 369–378.
https://doi.org/10.1177/0894439317710450.

Höhne, J.K., Schlosser, S., & Krebs, D. (2017). Inves-
tigating cognitive effort and response quality of
question formats in web surveys using paradata.
Field Methods, *29*(4), 365–382. https://doi.org/10.
1177/1525822X17710640.

Holland, J.L., & Christian, L.M. (2009). The influence of
topic interest and interactive probing on responses to
open-ended questions in web surveys. *Social Science
Computer Review*, *27*(2), 196–212. https://doi.org/
10.1177/0894439308327481.

Kaczmirek, L., & Neubarth, W. (2007). Nicht-reaktive Da-
tenerhebung: Teilnahmeverhalten bei Befragungen
mit Paradaten evaluieren. In DGOF (Ed.), *On-
line-Forschung 2007: Grundlagen und Fallstudien*
(pp. 293–311). Köln: Herbert von Halem Verlag.

Kaczmirek, L., Meitinger, K., & Behr, D. (2017). *Higher
data quality in web probing with EvalAnswer:
a tool for identifying and reducing nonresponse in
open-ended questions*. GESIS papers, Vol. 2017/1.
Mannheim: GESIS. https://doi.org/10.21241/ssoar.
51100.

Kunz, T., & Hadler, P. (2020). *Web paradata in survey research. GESIS Survey Guidelines*. Mannheim: GESIS—Leibniz Institute for the Social Sciences. https://doi.org/10.15465/gesis-sg_037.

Kunz, T., & Meitinger, K. (2022). A comparison of three designs for list-style open-ended questions in web surveys. *Field Methods*, *34*(4), 303–317. https://doi.org/10.1177/1525822X221115831.

Kunz, T., Beuthner, C., Hadler, P., Roßmann, J., & Schaurer, I. (2020). *Informing about web paradata collection and use. GESIS Survey Guidelines*. Mannheim: GESIS—Leibniz Institute for the Social Sciences. https://doi.org/10.15465/gesis-sg_en_036.

Kuusela, H., & Paul, P. (2000). A comparison of concurrent and retrospective verbal protocol analysis. *The American Journal of Psychology*, *113*(3), 387–404. https://doi.org/10.2307/1423365.

Lee, S., McClain, C. A., Webster, N., & Han, S. (2016). Question order sensitivity of subjective well-being measures: Focus on life satisfaction, self-rated health, and subjective life expectancy in survey instruments. *Quality of Life Research*, *25*(10), 2497–2510. https://doi.org/10.1007/s11136-016-1304-8.

Lee, S., McClain, C. A., Behr, D., & Meitinger, K. (2020). Exploring mental models behind self-rated health and subjective life expectancy through web probing. *Field Methods*, *32*(3), 309–326. https://doi.org/10.1177/1525822X20908575.

Lenzner, T., & Neuert, C. E. (2017). Pretesting survey questions via web probing—does it produce similar results to face-to-face cognitive interviewing? *Survey Practice*, *10*(4), 1–11. Retrieved from http://www.surveypractice.org/article/2768-pretesting-survey-questions-via-web-probing-does-it-produce-similar-results-to-face-to-face-cognitive-interviewing.

Lenzner, T., Kaczmirek, L., & Lenzner, A. (2010). Cognitive burden of survey questions and response times: a psycholinguistic experiment. *Applied Cognitive Psychology*, *24*(7), 1003–1020. https://doi.org/10.1002/acp.1602.

Lenzner, T., Neuert, C. E., & Otto, W. (2016). *Cognitive pretesting. GESIS survey guidelines*. Mannheim: GESIS—Leibniz Institute for the Social Sciences. https://doi.org/10.15465/gesis-sg_en_010.

Luebker, M. (2021). How much is a box? The hidden cost of adding an open-ended probe to an online survey. *methods, data, analyses*, *15*(1), 7–42. https://doi.org/10.12758/mda.2020.09.

Massen, C., & Bredenkamp, J. (2005). Die Wundt-Bühler-Kontroverse aus der Sicht der heutigen kognitiven Psychologie. *Zeitschrift für Psychologie*, *213*(2), 109–114. https://doi.org/10.1026/0044-3409.213.2.109.

Matjašič, M., Vehovar, V., & Lozar Manfreda, K. (2018). Web survey paradata on response time outliers: a systematic literature review. *Metodološki zvezki*, *15*(1), 23–41. https://doi.org/10.51936/yoqn3590.

Meitinger, K. (2017). Necessary but insufficient: Why measurement invariance tests need online probing as a complementary tool. *Public Opinion Quarterly*, *81*(2), 447–472. https://doi.org/10.1093/poq/nfx009.

Meitinger, K., & Behr, D. (2016). Comparing cognitive interviewing and online probing: do they find similar results? *Field Methods*, *28*(4), 363–380. https://doi.org/10.1177/1525822X15625866.

Meitinger, K., Braun, M., & Behr, D. (2018). Sequence matters in web probing: the impact of the order of probes on response quality, motivation of respondents, and answer content. *Survey Research Methods*, *12*(2), 103–120. https://doi.org/10.18148/srm/2018.v12i2.7219.

Meitinger, K., Behr, D., & Braun, M. (2019). Using apples and oranges to judge quality? Selection of appropriate cross-national indicators of response quality in open-ended questions. *Social Science Computer Review*, *39*(3), 434–455. https://doi.org/10.1177/0894439319859848.

Meitinger, K., Toroslu, A., Raiber, K., & Braun, M. (2022). Perceived burden, focus of attention, and the urge to justify: the impact of the number of screens and probe order on the response behavior of probing questions. *Journal of Survey Statistics and Methodology*, *10*(4), 923–944. https://doi.org/10.1093/jssam/smaa043.

Miller, A. L., & Lambert, A. D. (2014). Open-ended survey questions: item nonresponse nightmare or qualitative data dream? *Survey Practice*, *7*(5), 1–11. https://doi.org/10.29115/SP-2014-0024.

Miller, K., Willson, S., Chepp, V., & Padilla, J.-L. (Eds.). (2014). *Cognitive interviewing methodology*. Hoboken: John Wiley & Sons.

Naber, D., & Padilla, J.-L. (2022). *Nonresponse-related quality indicators of web probing responses and bias in cross-cultural web surveys*. Berlin: General Online Research (GOR).

Neuert, C. E., & Lenzner, T. (2021). Effects of the number of open-ended probing questions on response quality in cognitive online pretests. *Social Science Computer Review*, *39*(3), 456–468. https://doi.org/10.1177/0894439319866397.

Neuert, C. E., & Lenzner, T. (2023). Design of multiple open-ended probes in cognitive online pretests using web probing. *Survey Methods: Insights from the*

Field. https://doi.org/10.13094/SMIF-2023-00005. Advance online publication.

Neuert, C. E., Meitinger, K., & Behr, D. (2021). Open-ended versus closed probes: assessing different formats of web probing. *Sociological Methods & Research*, *35*(2), 1–35. https://doi.org/10.1177/00491241211031271.

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: verbal reports on mental processes. *Psychological Review*, *84*(3), 231–259. https://doi.org/10.1037/0033-295X.84.3.231.

Overgaard, M., & Sandberg, K. (2012). Kinds of access: different methods for report reveal different kinds of metacognitive access. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *367*(1594), 1287–1296. https://doi.org/10.1098/rstb.2011.0425.

Peytchev, A. (2009). Survey breakoff. *Public Opinion Quarterly*, *73*(1), 74–97. https://doi.org/10.1093/poq/nfp014.

Presser, S., Rothgeb, J. M., Couper, M. P., Lessler, J. T., Martin, E., Martin, J., et al. (2004). *Methods for testing and evaluating survey questionnaires*. Hoboken: John Wiley & Sons.

Priede, C., & Farrall, S. (2011). Comparing results from different styles of cognitive interviewing: 'verbal probing' vs. 'thinking aloud'. *International Journal of Social Research Methodology*, *14*(4), 271–287. https://doi.org/10.1080/13645579.2010.523187.

Reja, U., Lozar Manfreda, K., Hlebec, V., & Vehovar, V. (2003). Open-ended vs. closed-ended questions in web questionnaires. *Metodološki zvezki*, *19*, 159–177. Retrieved from http://mrvar.fdv.uni-lj.si/pub/mz/mz19/reja.pdf.

Revilla, M., & Couper, M. P. (2018). Comparing grids with vertical and horizontal item-by-item formats for PCs and smartphones. *Social Science Computer Review*, *36*(3), 349–368. https://doi.org/10.1177/0894439317715626.

Russo, J. E., Johnson, E. J., & Stephens, D. L. (1989). The validity of verbal protocols. *Memory & Cognition*, *17*(6), 759–769. https://doi.org/10.3758/BF03202637.

Scanlon, P. J. (2019). The effects of embedding closed-ended cognitive probes in a web survey on survey response. *Field Methods*, *31*(4), 328–343. https://doi.org/10.1177/1525822X19871546.

Scanlon, P. J. (2020). Using targeted embedded probes to quantify cognitive interviewing findings. In P. C. Beatty, D. Collins, L. Kaye, J.-L. Padilla, G. B. Willis & A. Wilmot (Eds.), *Advances in questionnaire design, development, evaluation and testing* (pp. 427–449). Hoboken: John Wiley & Sons.

Schmidt, K., Gummer, T., & Roßmann, J. (2020). Effects of respondent and survey characteristics on the response quality to an open-ended attitude question in web surveys. *methods, data, analyses*, *14*(1), 3–34. https://doi.org/10.12758/mda.2019.05.

Schober, M. F. (1999). Making sense of questions: an interactional approach. In M. G. Sirken, D. J. Herrmann, S. Schechter, N. Schwarz, J. M. Tanur & R. Tourangeau (Eds.), *Wiley series in probability and statistics survey methodology section. Cognition and survey research* (pp. 77–93). New York: John Wiley & Sons.

Schuman, H. (1966). The random probe: a technique for evaluating the validity of closed questions. *American Sociological Review*. https://doi.org/10.2307/2090907.

Schuman, H., & Presser, S. (1979). The open and closed question. *American Sociological Review*, *44*(5), 692–712. https://doi.org/10.2307/2094521.

Schwarz, N., Strack, F., Müller, G., & Chassein, B. (1988). The range of response alternatives may determine the meaning of the question: further evidence on informative functions of response alternatives. *Social Cognition*, *6*(2), 107–117. https://doi.org/10.1521/soco.1988.6.2.107.

Schwarz, N., Strack, F., & Mai, H.-P. (1991). Assimilation and contrast effects in part-whole question sequences: a conversational logic analysis. *Public Opinion Quarterly*, *55*(1), 3–23. https://doi.org/10.1086/269239.

Silber, H., Zuell, C., & Kuehnel, S.-M. (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. https://doi.org/10.5964/meth.2801.

Singer, E., & Couper, M. P. (2017). Some methodological uses of responses to open questions and other verbatim comments in quantitative surveys. *methods, data, analyses*, *11*(2), 115–134. https://doi.org/10.12758/mda.2017.01

Smith, T. W. (1989). Random probes of GSS questions. *International Journal of Public Opinion Research*, *1*(4), 305–325. https://doi.org/10.1093/ijpor/1.4.305.

Smyth, J. D., Dillman, D. A., Christian, L. M., & Mcbride, M. (2009). Open-ended questions in web surveys: can increasing the size of answer boxes and providing extra verbal instructions improve response quality? *Public Opinion Quarterly*, *73*(2), 325–337. https://doi.org/10.1093/poq/nfp029.

Sternberg, R. J. (1997). Construct validation of a triangular love scale. *European Journal of Social Psychology*, *27*(3), 313–335. https://doi.org/10.1002/(SICI)1099-

0992(199705)27:36lt;313::AID-EJSP824>3.0.CO; 2-4

Theofilou, P. (2013). Quality of life: definition and measurement. *Europe's Journal of Psychology*, *9*(1), 150–162. https://doi.org/10.5964/ejop.v9i1.337.

Tourangeau, R., Rips, L.J., & Rasinski, K.A. (Eds.). (2000). *The psychology of survey response*. Cambridge: Cambridge University Press.

Tourangeau, R., Conrad, F.G., Couper, M.P., & Ye, C. (2014). The effects of providing examples in survey questions. *Public Opinion Quarterly*, *78*(1), 100–125. https://doi.org/10.1093/poq/nft083.

Tukey, J.W. (1977). *Exploratory data analysis*. Reading: Addison-Wesley.

Veenhoven, R. (2000). The four qualities of life. *Journal of Happiness Studies*, *1*(1), 1–39. https://doi.org/10.1023/A:1010072010360.

Willis, G.B. (2005). *Cognitive interviewing: a tool for improving questionnaire design*. Thousand Oaks: SAGE.

Willis, G.B. (2015a). *Analysis of the cognitive interview in questionnaire design. Understanding qualitative research*. Oxford: Oxford University Press.

Willis, G.B., & Artino, A.R. (2013). What do our respondents think we're asking? Using cognitive interviewing to improve medical education surveys. *Journal of Graduate Medical Education*, *5*(3), 353–356. https://doi.org/10.4300/JGME-D-13-00154.1.

Wilson, T.D., Lafleur, S.J., & Anderson, D.E. (1996). The validity and consequences of verbal reports about attitudes. In N. Schwarz & S. Sudman (Eds.), *Answering questions: methodology for determining cognitive and communicative processes in survey research* (pp. 91–114). San Fransisco: Jossey-Bass.

Yan, T., & Tourangeau, R. (2008). Fast times and easy questions: the effects of age, experience and question complexity on web survey response times. *Applied Cognitive Psychology*, *22*(1), 51–68. https://doi.org/10.1002/acp.1331.

Yan, T., & Williams, D. (2022). Response burden—review and conceptual framework. *Journal of Official Statistics*, *38*(4), 939–961. https://doi.org/10.2478/jos-2022-0041.

Yan, T., Fricker, S., & Tsai, S. (2020). Response burden: What is it and what predicts it? In P.C. Beatty, D. Collins, L. Kaye, J.-L. Padilla, G.B. Willis & A. Wilmot (Eds.), *Advances in questionnaire design, development, evaluation and testing* (pp. 193–212). Hoboken: John Wiley & Sons.

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

Zuell, C., Menold, N., & Körber, S. (2015). The influence of the answer box size on item nonresponse to open-ended questions in a web survey. *Social Science Computer Review*, *33*(1), 115–122. https://doi.org/10.1177/0894439314528091.