Universität Mannheim

Fakultät für Sozialwissenschaften

CONTEXT EFFECTS IN QUESTION EVALUATION VIA WEB PROBING: EXPLORING THE INTERACTION OF OPEN-ENDED AND CLOSED SURVEY QUESTIONS

Inauguraldissertation
zur Erlangung des akademischen Grades
einer Doktorin der Sozialwissenschaften
der Universität Mannheim

Vorgelegt von J. Patricia Hadler, M.A.

Hauptamtlicher Dekan der Fakultät für Sozialwissenschaften:
Prof. Dr. Michael Diehl

Erstbetreuerin und Erstgutachterin:
Prof. Dr. Natalja Menold

Zweitbetreuer und Zweitgutachter:
Prof. Dr. Herbert Bless

Drittgutachter:
Prof. José-Luis Padilla García, PhD

Tag der Disputation: 30. November 2023

ACKNOWLEDMENTS

Writing and completing a thesis is only possible with the support of many people. Foremost, I would like to thank my doctoral supervisor Prof. Dr. Natalja Menold, who opened the door for me to return to academia and motivated me to pursue research in survey methodology. Her critical advice and encouraging feedback are the true foundation of this thesis. Moreover, I would also like to express my gratitude to my second supervisor Prof. Dr. Herbert Bless and my third reviewer Prof. Dr. José-Luis Padilla, for immediately being on board and for their time and effort as members of the dissertation committee.

I want to express my gratitude for the pretesting team at GESIS, namely Dr. Timo Lenzner, Dr. Cornelia Neuert and Lukas Schick. Working in the pretesting unit sparked my enthusiasm for question evaluation methods, and the team spirit and my colleagues' continuous efforts made it possible for me to find the time to work on this thesis. Thank you so much.

My empirical studies were embedded in cooperative research projects at GESIS. I would like to thank Prof. Dr. Beatrice Rammstedt for providing the means to carrying out these studies, and my colleagues who shared the responsibility for the technical and organizational implementation, in particular Dr. Ranjit Singh, Dr. Verena Ortmanns, Dr. Cornelia Neuert and Dr. Timo Lenzner. I am grateful to the support I received form Dörte Naber, Patricia Steins, Friederike Quint, and Marisa Finzer in analysing the data.

I greatly benefited from the encouraging and friendly environment at GESIS. Next to my immediate team, many thanks go to Dr. Dorothée Behr, Dr. Matthias Blümke, Prof. Dr. Michael Braun, Dr. Brita Dorer, Ulrike Efu Nkong, Dr. Tanja Kunz, Prof. Dr. Katharina Meitinger, Wanda Otto, Dr. Lydia Repke, Matthias Roth, Dr. Matthias Sand, Dr. Ines Schaurer, Dr. Silke Schneider, Cornelia Züll, and all my current and former colleagues at the institute who are never too busy to find a kind word.

Finally, I thank my family for their unwavering love and believing. To my mother and father for routing me on for as long as I can remember and longer. To my brothers Bobby and Michael and their families for their support and friendship. And to my family in-law through whom I have been blessed with additional parents, brothers, and sisters. In particular, I thank my husband Jonathan for his love and companionship not just during this thesis, but throughout every step I have taken in the past seventeen years. Finally, I am deeply grateful for Jakob and Johan for being in our lives. Throughout countless projects and deadlines, you have taught us never to lose sight of what is truly important.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

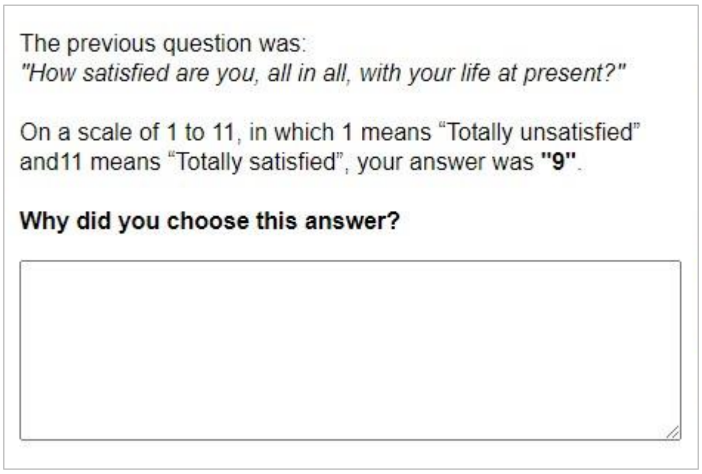
1. GENERAL INTRODUCTION

Surveys collect data on respondents' beliefs, values, attitudes, behaviours, and states of affairs by asking *questions* (Bradburn, Sudman, & Wansink, 2004) and collecting respondents’ *answers* (Schwarz & Sudman, 1996). At their core, surveys constitute an—albeit unusual—form of communication, perhaps best described as an indirect dialogue between an absent researcher and a respondent (Conrad, Schober, & Schwarz, 2014; Schwarz, 1995). Even when surveys are carried out via an interviewer, this intermediary is generally instructed not to provide additional information to the respondent that goes beyond scripted clarifications (Schober, 1999). Therefore, it is essential that respondents can independently comprehend a survey question, retrieve the relevant information, form a judgment and communicate their internal response using the available response format (Tourangeau, Rips, & Rasinski, 2000). Respondents’ interpretation of survey questions and, ultimately, their responses do not only depend on question wording (Lenzner, 2011; Schaeffer & Dykema, 2020) but also on contextual elements, such as visual design (Couper, Conrad, & Tourangeau, 2007; Toepoel & Couper, 2011; Toepoel & Dillman, 2011) or preceding survey questions (Rasinski, Lee, & Krishnamurty, 2012; Smyth, Dillman, & Christian, 2007; Stark et al., 2020). To ensure data quality (Rammstedt et al., 2015), researchers must verify that respondents understand and respond to survey questions in the way intended by means of question evaluation and testing (Beatty et al., 2020; Presser et al., 2004).

1.1. Probing as a method of question evaluation

Various methodological approaches to question evaluation exist (Beatty et al., 2020; Groves et al., 2011). Cognitive pretesting is a qualitative approach that seeks to uncover respondents’ mental processes while answering survey questions (Miller, Willson, Chepp, & Padilla, 2014). In cognitive interviews, respondents may be asked to *think aloud* while answering a question, or they are asked so-called probing questions or simply *probes* after having responded to the survey question (Collins, 2015; Padilla & Leighton, 2017; Priede & Farrall, 2011). Examples of typical probes are “How did you arrive at that answer?” or “What does the term XY mean to you in this question?” (see Willis, 1994, for an overview of probing techniques). Asking single probes was first promoted in the context of production surveys as “embedded probing” (Converse & Presser, 1986) and administered to a subsample of the survey population, indicated by the term “random probe” (Converse & Presser, 1986; Schuman, 1966; Smith, 1989). Today, in-depth cognitive interviews predominantly take place in a debriefed setting during questionnaire development and pretesting, often with trained interviewers in cognitive pretesting laboratories (Sirken et al., 1999). The data collected through cognitive interviews is analysed qualitatively (Willis, 2015a) to determine whether the survey question generates the information the researcher intended (Beatty & Willis, 2007). This may be done following a descriptive approach that examines how respondents construct the pragmatic meaning of a survey question (Miller, 2014; Padilla & Benítez Baena, 2014), or following a reparative approach, primarily focussing on identifying and resolving problems respondents encounter while responding to the survey question (Meadows, 2021; Willis, 2015a).

Mirroring the rise of web surveys as a self-administered mode of survey data collection (Groves, 2011; Leeuw, 2018), web probing has evolved as a self-administered method of question evaluation that implements probing techniques from cognitive interviewing in web surveys (Behr, Meitinger, Braun, & Kaczmirek, 2017). In a typical web probing setting, a respondent first answers a closed survey question. On the next survey page, the respondent is presented the probing question, for instance asking her to explain why she selected a particular response option (see Figure 1.1).



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

On a scale of 1 to 11, in which 1 means “Totally unsatisfied” and11 means “Totally satisfied”, your answer was **"9"**.

Why did you choose this answer?

Figure 1.1. Example of web probing

Due to its comparatively simple and standardized implementation, web probing has become popular in the context of evaluating questions in cross-national surveys (Behr, Meitinger, Braun, & Kaczmirek, 2020) to gain insights on reasons for lacking measurement invariance (Leitgöb et al., 2022; Meitinger, 2017) and to evaluate web survey questions in the same mode (Fowler & Willis, 2020). Web probes can be used at the stage of questionnaire development and pretesting, but also be implemented during a production web survey, or even in post-hoc evaluations to gain insights on unexpected survey results (i.e., Behr, Braun, Kaczmirek, & Bandilla, 2014). Web probe responses are analysed qualitatively by applying inductive or deductive coding schemes (Willis, 2015a). Due to the higher case numbers as compared to cognitive interviewing, these codes are increasingly employed in subsequent quantitative analyses, such as subgroup comparisons (Neuert, Meitinger, & Behr, 2021) or to explain survey response behaviour (Behr, Braun et al., 2014; Meitinger, 2018). Web probing lacks the cognitive interviewer who can motivate respondents to answer probes or follow up on responses that remain ambiguous (Meitinger & Behr, 2016) and requires both general and computer literacy (Callegaro, Lozar Manfreda, & Vehovar, 2015, p. 65), as respondents must independently type in their responses to the probing questions (Meitinger & Behr, 2016). In summary, although the techniques used in web probing are not new, transferring probes used in cognitive interviews into web surveys has several implications for their implementation and the data collected in this way. However, regardless of mode, the purpose of probes is to examine how respondents construct the meaning of a survey question and whether this coincides with the researcher's intention.

1.2. Context effects and probing

Context effects remind us that the interpretation of and response to a survey question is influenced by more than the question's wording (Smyth et al., 2007). The term describes situations where an identical survey question "produces different answers depending on the context in which it is asked" (Tourangeau, 1999, p. 111). The most prominent types of context effects are question order effects. To name just one classic example, the correlation between life and relationship satisfaction depends on the order of the survey questions, whether respondents perceive these questions as pertaining to the same overarching topic, and ultimately infer that they should include or exclude their relationship satisfaction in the evaluation of their overall life satisfaction (Schwarz, Strack, & Mai, 1991; Tourangeau, Rasinski, & Bradburn, 1991).

Context effects result from the cognitive processes underlying question construal (Bless & Schwarz, 2010; Tourangeau et al., 2000) and the application of communicative principles (Clark & Haviland, 1977; Conrad et al., 2014; Grice, 1975). In a rough summary of previous research, context effects are likely to occur when questions are perceived as standing in relationship to each other (Strack, 1992) and directly follow each other (Tourangeau, Singer, & Presser, 2003), especially when the question topic of the target question is unfamiliar or worded ambiguously (Tourangeau, 1999).

Despite general agreement that questions in surveys are always understood and answered in light of the context they are asked in, methodological research on cognitive probing in surveys has—to a large extent—implicitly continued to assume that respondents answer survey questions and probes independently of surrounding questions. This is surprising as settings that include probes are likely to include factors that contribute to the emergence of context effects. By nature, probes directly relate to the survey questions they pertain to, establishing a close connection between the survey question and probe (Silber, Zuell, & Kuehnel, 2020). Moreover, probes are likely to be implemented when a term used in a survey question may be unclear to respondents. Thus, it seems plausible that respondents rely on contextual elements such as preceding questions when responding to probes or that probing questions contribute to the context in which survey questions are construed. Unfortunately, theoretical discussions and empirical studies on context effects involving probes are lacking (though see Conrad & Blair, 2009, for a rare discussion in the context of cognitive interviews).

Context effects that arise in survey settings that include probes are relevant from two perspectives. First, the context in which a probe is asked may impact the data collected by this probe. For instance, the insights gained through probing may vary depending on question order, resulting in a different evaluation of what a survey question under examination is truly measuring (descriptive approach) and whether it requires revision (reparative approach). Understanding such effects is essential when probing is used during question development and pretesting. Secondly, implementing probes may impact the data collected by the survey questions, resulting in other survey estimates. While such effects may be deemed of secondary importance in the context of question pretesting, they have wide-reaching implications if probes are implemented in production surveys, as has been suggested by multiple researchers, especially for self-administered web surveys (i.e., Meitinger, 2017; Singer & Couper, 2017).

1.3. Aims and structure of the thesis

This thesis aims to develop a theoretical and empirical understanding of context effects in web probing. It does this by (1) establishing a psychological model of context effects when survey questions are evaluated using web probes and (2) examining this model in a series of empirical studies. The focus lies on the effects of the sequence of survey questions and probes on survey and probe responses. The insights from the thesis contribute to several important research fields in survey methodology. For one, they are relevant to researchers implementing probes for question evaluation. More generally, they shed light on the interaction of open-ended and closed questions and enhance our understanding of question order effects in surveys.

The basic premises underlying this thesis are that answering survey questions requires several cognitive steps (Tourangeau et al., 2000) and that these steps are impacted by question context, for instance question order. The task of probing adds complexity and dynamics to this question-answer process and the context effects that potentially accompany it. Probes are not simply questions on the same topic as a survey question, but on the survey question itself. Answering probes requires the metacognitive tasks of introspection and retrospection (Ericsson & Simon, 1980). Introspection is known to cause reactivity and retrospection is prone to memory errors and post-rationalization (Bröder, 2019; Massen & Bredenkamp, 2005). This results in a wide range of potential context effects when asking respondents to reflect on their thought processes while answering survey questions. In short, survey questions may impact probe responses and probes may impact survey responses. Finally, when several probes are asked within one survey, these probes may impact each other due to the Gricean maxims of communication (Grice, 1975), such as the norm of non-redundancy (Clark & Haviland, 1977).

The remainder of this thesis is arranged as follows. Chapter 2 introduces web probing as a method of question evaluation and as the object of investigation in this thesis. It begins by tracing the development of web probing from its origins in cognitive interviewing, focusing on the impact of mode on implementing probes and the data collected in this way. Next, it provides an overview of the standard practices of implementing probing techniques in web surveys. Finally, the chapter summarizes for which analytical purposes web probing is currently employed. Chapter 3 reviews the role of cognitive and communicative processes in the emergence of context effects and lays the theoretical foundation to establish a framework for context effects in web probing in the subsequent chapter. It does this by illustrating what context effects are, how they can be classified, and which cognitive models have been used to explain them, such as the cognitive process of survey response (Tourangeau et al., 2000) and the inclusion/exclusion model (Bless & Schwarz, 2010) as well as the communicative maxims (Clark & Haviland, 1977; Grice, 1975) that are integrated into these models. Next, the cognitive processes underlying context effects are illustrated for specific survey settings relevant to web probing. These are web surveys as a survey mode, response burden as a function of question context, and open-ended and closed questions as question formats. The final part of Chapter 3 describes the cognitive process of probing as a task that relies on introspection and retrospection and how this impacts the depth of survey question processing. The chapter closes with an excursus in how far probing is an effective method of examining thought processes.

Chapter 4 combines the knowledge on context effects in surveys with insights on the nature of cognitive probing to establish a psychological model of context effects in web probing. It begins by distinguishing the directions of context effects in web probing and classifying the range of possible research designs, as these are more diverse and complex than those used in research on survey questions only. Based on this classification, a model of context effects in web probing is introduced. The respective subchapters elaborate on the effects of survey questions on probe responses, the effects of probes on survey responses, and the effects of probes on responses to other probes. Each section summarizes previous research and points to existing research gaps. The final subchapter derives the research questions from the research gaps and closes with an overview of how the empirical studies address the research questions.

The first study in Chapter 5 examines whether intermittent survey questions increase response burden and memory errors for probe responses that rely on retrospection. To this end, it randomizes whether probes are asked directly after the survey question they pertain to, or later in the survey. In addition, the study examines whether these effects differ depending on whether probes are asked in the more common—but cognitively demanding—open-ended narrative format or using predefined response options. The second study in Chapter 6 is based on the notion of reactivity through introspection and examines the effects of probes on surrounding survey questions. It distinguishes between effects on preceding and subsequent questions and effects on survey break-off. The third study in Chapter 7 examines the effect of the sequence of the survey questions (and ultimately probes) on both survey and probe responses. It analyses how changing the order of the survey questions impacts the consistency of probe responses to attitude and behaviour questions and whether respondents are reluctant to reiterate content they have already shared in a previous probe response. Chapter 8 summarizes the main findings, discusses how they support the model established in Chapter 4 or merit changes to it, and derives practical recommendations for researchers employing cognitive probes for question evaluation. It closes by suggesting directions for future research to advance our understanding of context effects in question evaluation.

2. WEB PROBING AS A METHOD OF QUESTION EVALUATION

This chapter introduces web probing as a method of evaluating survey questions. It begins with the origins of web probing in cognitive interviewing and which implications the switch to self-administration has for researchers employing probes. Next, it gives an overview of the standard practices of implementing probing techniques in web surveys. The chapter concludes with typical settings in which web probing is utilized.

A probe is a follow-up question to a preceding survey question. Probing techniques range from general, non-directive questions, such as “Could you tell me more about that?”, to rather specific, directive probes that ask about a particular aspect of the question-answer process, such as “Which groups of immigrants did you have in mind when you answered the question?” (see Foddy, 1998, for a discussion of the directiveness of probes). The goal of cognitive probing is “to evaluate the quality of the response or to help determine whether the question is generating the information that its author intends” (Beatty & Willis, 2007, p. 288). Web probing is “the implementation of probing techniques from cognitive interviewing in web surveys” (Behr et al., 2017, p. 1).

2.1. From cognitive interviewing to web probing

Web probing emerged in the 2000s as an online extension of and supplement to probe-based cognitive interviewing (Behr et al., 2017). In contrast to simultaneous efforts to carry out personal cognitive interviews via online video platforms (i.e., Mockovak & Kaplan, 2015), this development was marked by an effort to collect probe responses in self-administered settings. Braun (2008) inserted written probes into a web survey and asked respondents to type in their response in an early example that strongly resembles today’s standard practice. Other researchers experimented with voice-over-Internet protocols that read out probes to respondents (Mohorko & Hlebec, 2016), audio-recording respondents’ spoken answers to probes that were presented in writing (Edgar, Murphy, & Keating, 2016; Murphy, Keating, & Edgar, 2013), and instant messaging (Mohorko & Hlebec, 2016) (see Yu et al., 2019 for an overview of web-based cognitive pretesting methods used by the U.S. Census Bureau). The development of the current common practice of web probing with written probes and typed responses was accelerated by a research project on improving intercultural comparative research (German Research Foundation, SPP 1292, project 161767778) (Behr, Braun et al., 2014, 2012; Behr, Kaczmirek, Bandilla, & Braun, 2012b; Braun, Behr, & Kaczmirek, 2013).

Today, cognitive interviewing and web probing remain the two most common practices within the probing paradigm and represent opposing poles regarding interviewer involvement. On the one extreme, cognitive interviewing requires the presence of an interviewer, asks probes orally, and collects data in the form of verbal responses (Conrad, Blair, & Tracy, 1999). On the other extreme, web probing takes place in a self-administered setting without an interviewer, presents probes in written form, and collects answers typed by the respondent (Behr et al., 2020; Fowler & Willis, 2020). The implications of this are manifold for both the data collection process and, ultimately, the data collected via probes in cognitive interviews and web surveys.

Probe-based cognitive interviews are usually conducted as semi-scripted interviews (Collins, 2015), meaning that interviewers employ a scripted interview protocol but are trained to react flexibly to the interview situation and deviate from the script when required, for instance, to motivate respondents or to ask follow-up questions when a respondent's initial answer remains unclear (Willis, 1994, 2005). In contrast to the flexibility of cognitive interviews, web probing is restricted to scripted probes. There is no interviewer to motivate respondents or clarify ambiguous answers, resulting in higher levels of uninterpretable answers and probe nonresponse (Meitinger & Behr, 2016).

While cognitive interviews do not suffer from high levels of survey or probe nonresponse, they often suffer in terms of comparability across interviews, interviewers, institutions, countries, languages, or cultures (Conrad & Blair, 2009; Priede, Jokinen, Ruuskanen, & Farrall, 2014; Rothgeb, Willis, & Forsyth, 2007; Willis, 2015b). The restriction to scripted probes in web probing enables a much higher level of standardization, with no issues of comparability across interviews or—professional translation provided—even across languages and countries (Behr et al., 2020).

As qualitative methods, neither cognitive interviews nor web probing are designed to be representative of a larger population. In cognitive interviews, sample composition may be solely guided by theoretical saturation (Beatty & Willis, 2007; Blair & Conrad, 2011; Padilla & Benítez Baena, 2014), though most cognitive interview studies employ demographic quotas. Recruitment strategies for cognitive interviews vary strongly. There are rare cases of probability sampling (i.e., Oksenberg, Cannell, & Kalton, 1991); however, the norm is convenience sampling, relying on snowball recruitment, classified or even social media advertising (Head, Dean, Flanigan, Swicegood, & Keating, 2016). Often, cognitive interviews are carried out within one or a few geographical regions near a cognitive pretesting institution (Behr et al., 2017).[1] Cognitive interviews are often limited to small sample sizes, potentially failing to detect problems with the question-answer process (Blair & Conrad, 2011). Web probing can be implemented quickly, without the need to train interviewers, and can collect data from large samples in a short time. The target population in web probing underlies the same restrictions as web surveys, being limited to people with internet access and sufficient literacy, in particular as open-ended probes require not only sufficient reading, but also writing skills (Couper, 2000; Galesic, 2006; Galesic & Bosnjak, 2009). Though some recent web probing studies have employed probability-based web panels (i.e., Irimata & Scanlon, 2022; Willson, Scanlon, & Miller, 2022), recruitment is typically carried out using commercial panels with self-selection, such as opt-in online access panels, crowdsourcing or social media platforms (Behr et al., 2012b; Edgar et al., 2016).

Both cognitive interviews and web probing are regularly used to pretest or evaluate questions used in general population surveys. Moreover, cognitive interviews are popular in questionnaire evaluation for special target groups such as illiterate people, the elderly, or people with specific illnesses (see Drennan, 2003, on the use of cognitive interviews in health care research; see Jobe & Mingay, 1990, for cognitive interviewing with the elderly). Web probing is becoming increasingly popular in the context of cross-national surveys, as cross-cultural cognitive interview projects are logistically complex (Willis, 2015b, p. 360) and highly resource-intensive (see Fitzgerald, Widdop, Gray, & Collins, 2011 for an example of cross-cultural cognitive interviewing). Web probing enables elaborate sampling plans based on demographics or responses to survey questions under evaluation (Behr et al., 2017, 2020).

While cognitive interviews are primarily implemented at the stage of questionnaire development and pretesting (for a list of exceptions, see Behr, Braun et al.,

[1] This is not to ignore that remote cognitive interviewing has increased in recent years (for an early example, see Mockovak & Kaplan, 2015), partially amending the issue of regional diversification (for cognitive interviews via telephone, see Noel, 2013; for cognitive interviews via video software, see Fry, Mitchell, & Wiener, 2021). Schober et al. (2020) provide for an overview of aspects to consider when switching to remote interviewing.

2014), researchers have advocated for applying web probing at all stages of the web survey life cycle, ranging from pretesting draft questionnaires (Fowler & Willis, 2020), to the implementation of probes into production web surveys (Behr et al., 2012b; Scanlon, 2019; Singer & Couper, 2017), to employing probes in post-hoc evaluations, for instance to explain unexpected survey response behaviour (Behr, Braun et al., 2012) or low survey data quality (Meitinger, 2017).

Given the complementary benefits and challenges of cognitive interviewing and web probing, researchers have suggested various combinations of the methods. For instance, some researchers have advocated conducting cognitive interviews and quantifying the findings using web probes (Behr et al., 2012b; Scanlon, 2020; Willson, Scanlon et al., 2022). Other studies have carried out initial cognitive pretesting using web probing and followed up on open research questions using in-depth cognitive interviews, including niche populations that could not be well reached online (Hadler, Neuert, Lenzner, & Menold, 2018) or used both methods simultaneously to examine different research questions (Hadler, Lenzner, Schick, & Neuert, 2022). Other conceivable combinations include carrying out cognitive interviews in single countries and complementing this by web probing studies in a larger number of countries or carrying out cognitive interviews during questionnaire development and using web probing in post-hoc studies to follow up on unexpected survey results.[2] The following section describes how web probing is implemented.

2.2. Web probing techniques and design

Web probing generally employs directive probes, the most common probing techniques being category selection, comprehension and specific probing (Behr et al., 2020). A *category selection probe* requests that respondents explain why they chose a specific response to a survey question. For instance, respondents might answer a closed survey

[2] Moreover, researchers advocate for combining different methods of survey question evaluation, in particular cognitive pretesting methods with methods such as expert reviews, appraisal-based methods, psychometric results, or latent class analysis. For studies arguing to combine cognitive pretesting with other question evaluation methods, see Behr et al., 2020; Benítez Baena & Padilla, 2014. For general recommendations of combining question evaluation methods, see Maitland & Presser, 2016, 2018; Yan, Kreuter, & Tourangeau, 2012.

question on their general health ("How would you rate your overall health?") with answer options ranging from "very good" to "very poor" (Bruin, Picavet, & Nossikov, 1996). A category selection probe might be worded as follows: "Please explain why you rated your health as 'poor'". A *comprehension probe* asks respondents to explain how they understand a certain term or phrase in the context of the survey question. A comprehension probe about the same question could be, "What does the term 'health' mean to you in this question?". A *specific probe* requires details on a particular aspect of the survey question. For instance, for a survey question on life satisfaction ("How satisfied are you at present, all in all, with your life?" (Beierlein, Kovaleva, László, Kemper, & Rammstedt, 2014; Schwarz, Strack, & Mai, 1991), a specific probe might ask respondents to enumerate which aspects of their life they considered in their answer ("Which aspects of your life did you consider when answering the question?").

Probes in web surveys are usually implemented as open-ended narrative questions, employing a paging design and concurrent probing (see Table 2.1 for an overview of key aspects of probe implementation). In a paging design, a respondent answers the survey question on one survey page and receives the probe on a separate survey page. The rationale is "to disentangle the response process for the closed-ended questions from the probing process and thus to keep the 'usual' survey experience of closed-ended questions as stable as possible" (Behr et al., 2017, p. 6). It is also possible to present a probe alongside the survey question on the same page (referred to here as embedded probe presentation; Neuert & Lenzner, 2021).[3] Embedding probes on the same page as the survey question has been done in the context of sensitive questions to allow respondents to explain their responses and prevent social desirability bias or item nonresponse (Couper, 2013; Luebker, 2021). A survey question may be followed by only one or several probes, which may be presented together on one survey page or with each probe on a separate page (Meitinger, Braun, & Behr, 2018; Meitinger, Toroslu, Raiber, & Braun, 2022).

[3] Please note that the term embedded probing has also been used to refer to concurrent probe placement and to the implementation of probes in production surveys (Scanlon, 2016, 2019, 2020).

Table 2.1. Key aspects of probe implementation

Design element	**Types**	**Definition**
Probe wording		
Directiveness	General	Request to type any thoughts regarding the survey question
	Directive	Request pertains to a specific aspect of the survey question or mental processing (see Techniques)
Technique *(selection)*	Category selection probe	Request to explain why one chose a particular response category
	Comprehension probe	Request to define a term within the question context
	Specific probe	Request to give details on a specific aspect of a question
Probe presentation		
Paging versus embedded	Paging	Survey question and probe are presented on separate survey pages
	Embedded	Survey question and probe are presented on the same survey page
Placement *(paging design only)*	Concurrent	Directly after the survey question
	Retrospective	After a block of survey questions or at the end of a survey
Only for multiple probes pertaining to one question:		
Sequence of probes	-	Order of presenting probes when multiple probes refer to one survey question
Number of probes per survey page	One	One probe per survey page
	Two or more	Several/all probes relating to one question on one survey page
Probe format		
Open-ended versus closed	Open-ended	Text field to type in response without predefined response options
	Closed / Semi-open	Predefined (check-all-that-apply or single-choice) response options
Text box design *(open-ended only)*	Multi-line	Large text box for narrative text
	List-style	One or several one-line text boxes; dynamic number of text boxes possible

Probes can be placed directly following a survey question (concurrent probe placement) or later in a survey, following a block of intermittent survey questions or at the end of a survey (retrospective probe placement) (Fowler & Willis, 2020). Concurrent probing follows the rationale that the respondents' thought processes are still available in short-term memory. Retrospective probing has been advocated to prevent probes from interfering with other survey questions (Drennan, 2003; Willis, 2005). Regardless of probe placement, both the survey question and, when relevant, the respondent's answer to the survey question are repeated above the probe (Behr et al., 2012b).

Probes are usually implemented as open-ended narrative questions with multi-line answer boxes, so that respondents can freely type in their responses. However, in recent years researchers and practitioners have tested the use of other open-ended formats (for instance several one-line answer boxes for probes requiring list-style answers, Meitinger & Kunz, 2022) and the implementation of semi-open and closed probes with predefined response options to lower response burden and facilitate analysis (Scanlon, 2018, 2020).

Finally, researchers must decide on the number of probes they wish to implement in a web survey, and whether to inform respondents at the beginning of the survey to expect questions about the survey questions.

2.3. Settings for employing web probing

Web probing aims to assess how a survey item is understood and whether respondents encounter difficulties during their response process. Consequently, the focus lies more on a single question than a measurement instrument or the questionnaire as a whole.

In many studies, probe responses are coded qualitatively to assess whether the respondents' understanding of a question is in scope with the underlying construct. For instance, Hadler, Neuert, Ortmanns, and Stiegler (2022) found that most respondents demonstrated an in-scope, albeit vague, understanding of a newly implemented non-binary sex category in Germany. In contrast, web probing of the general national pride item revealed that the question was associated with various aspects of national identity, of which a substantive portion did not align with the underlying construct (Meitinger, 2018). In a study on public health, web probes identified that almost all respondents well understood a newly developed question on COVID-19 testing, whereas a question about whether the respondents' health provider offered telemedicine access was misinterpreted

(Irimata & Scanlon, 2022; Willson, Scanlon et al., 2022). Other web probing studies have compared item comprehension across subgroups, for instance regarding the interpretation of job quality indicators by employees and self-employed (Hadler et al., 2018; Hadler, Lenzner et al., 2022). Some studies have used the qualitatively coded probe responses in subsequent quantitative analysis to explain survey response behaviour. For instance, a code indicating a low level of trust in the government strongly predicted agreeing with an item on civil disobedience (Behr, Braun et al., 2014).

Due to its comparatively simple implementation in cross-national web surveys, web probing has become a popular means of assessing the cross-cultural comparability of items. For multi-item inventories, web probing has been used in post-hoc analyses to determine reasons for the lack of measurement invariance (Leitgöb et al., 2022; Meitinger, 2017). For single-item measures that cannot be subjected to tests of measurement invariance using multi-group confirmatory factor analysis (MGCFA), web probing can provide insights into cross-cultural comparability (Meitinger, 2018). For instance, regarding the item above on civil disobedience, cross-cultural differences in the associations with the term explained differences in the levels of agreement across countries (Behr, Braun et al., 2014). Moreover, the already diverse associations with the item of general national pride differed across countries (Meitinger, 2018). Similarly, an examination of a single-item measure of cosmopolitanism found that respondents' understanding of what it means to be "citizens of the world" did not always align with the construct and cross-country differences existed as to why respondents agreed or disagreed with the item (Braun, Behr, & Díez Medrano, 2018).

Web probing has been implemented in split-ballot experiments with different question drafts, for instance to determine whether the response selection can be unambiguously attributed to an underlying attitude in both question versions (Braun, Meitinger, & Behr, 2020) or which of several possible terms best captures a construct such as "conflict" (Schick, Lenzner, Hadler, & Neuert, 2023).

Finally, web probing has proven helpful in examining response patterns across multiple survey questions. For instance, probe responses have been used to quantify the aspects considered in measures of self-rated health as compared to subjective life expectancy (Lee, McClain, Behr, & Meitinger, 2020). In other cases, the reasons behind seemingly contradictory survey responses have been examined. For instance, Braun and Johnson (2018) examined survey response patterns to items on xenophobia via probes, finding that the items did not form a unidimensional scale. In another study, reasons for seemingly contradictory responses to traditional and egalitarian gender items were examined by comparing probe responses across different survey response combinations (Behr, Braun et al., 2012).

In summary, a single probe often relates to a single survey question or item. However, web probing analysis may well extend beyond the scope of a single item, for instance to contribute to validity evidence of a multi-item measure (i.e., Behr, Braun et al., 2012; Meitinger, 2017) or to examine how respondents understand and respond to survey questions in relation to one another (i.e., Braun & Johnson, 2018; Lee et al., 2020). In these examples, a connection between several survey questions or items is assumed, or—in the case of multi-item measures that capture a latent construct—even required. When survey questions are deemed to be part of an overarching topic and are deemed by the respondent to be part of one communicative setting, this is the moment in which the possibility of context effects must be considered. The following section is dedicated to such context effects in surveys and their underlying cognitive and communicative processes.

3. COGNITION, COMMUNICATION AND CONTEXT EFFECTS

This chapter provides a theoretical basis which context effects may be expected in web probing by giving an overview of the cognitive and communicative processes that underly answering survey and probing questions and how they contribute to the emergence of context effects in surveys. The chapter begins by reviewing how context effects are defined and which types can be distinguished. Next, it summarizes cognition- and communication-based models that have been employed to explain the emergence of context effects, including the cognitive model of survey response (Tourangeau et al., 2000), the inclusion/exclusion model (Bless & Schwarz, 2010; Schwarz & Bless, 2007) and Grice's maxims of communication (1975). The third subchapter highlights specific survey aspects and settings relevant to web probing and how context effects emerge in them. It discusses the mode of web surveys, response burden as a dynamic feature and consequence of survey context, and open-ended and closed question formats. The final section of this chapter switches the focus from survey questions to probes. Unlike survey questions, probing poses a metacognitive task by asking questions *about* questions (Tanur, 1992). This requires that respondents carry out introspection and retrospection. The chapter closes with a discussion of the efficacy of probing in generating insights into respondents' mental processes.

"Context effects" is an umbrella term that summarizes influences on survey response behaviour due to the survey environment (Smyth et al., 2007). Often, context effects are equated with question order effects, that is, the "effects of earlier questions on answers to later ones" (Tourangeau et al., 2000, p. 200), and have also been coined spill-over, carry-over, or backfire effects (Dillman, Smyth, & Christian, 2014, p. 234-241; Schwarz, Knäuper, Oyserman, & Stich, 2008). However, context effects also encompass the effects of the order of response options (Garbarski, Schaeffer, & Dykema, 2015), the visual presentation of a question (Couper et al., 2007; Couper, Tourangeau, & Kenyon, 2004), and potentially even broader aspects such as a survey's title, stated purpose or sponsor (Galesic & Tourangeau, 2007), the presence or absence of other people (Tourangeau & Yan, 2007), the weather or respondents' mood at the time of a survey (Schwarz & Strack, 1999).

The focus of this chapter and the dissertation as such is question order effects, that is the effects of the sequence of survey and probing questions on survey and probe responses. This is because the sequence of asking survey questions and probes is central to web probing design and the presence of two types of questions (survey questions and probes) offers a complex array of potential question sequences. For instance, implementing a probe concurrently means that survey questions do not directly follow each other. In contrast, retrospective placement implies that the probe does not directly follow the survey question. These examples do not even include the classical setting for examining question order effects, in which the order of two related survey questions is varied, and how probes about these questions may be presented. However, it should be acknowledged that context effects unrelated to question order may also occur in web probing. For example, framing a web probing study as a cognitive online pretest compared to integrating probes in an unannounced production environment may impact respondents' likelihood of participating or completing a questionnaire. The effects of response option order will not be central to this overview, as they are seldom the focus of web probing. Issues of web survey programming and visual design will be discussed whenever they may contribute to the emergence of question order effects, as respondents treat formal features of a questionnaire as potentially relevant contributions of the absent researcher to the "survey conversation" (Schwarz, 1996). Because the effects of question order are ultimately the effect of the cognitive and communicative context in which a respondent processes a question, the term "context effects" will be used except when distinguishing between question order effects and other context effects.

3.1. Describing context effects in surveys

The simplest method of classifying context effects is based on describing the response behaviour to the target item in relation to the preceding or contextual item (Rasinski et al., 2012). When the response to the target question becomes more consistent or similar to the preceding question, this is known as an assimilation effect. When a preceding question leads to respondents giving a less consistent or similar answer, this is called a contrast effect (Tourangeau, 1999). Changes in response behaviour can be described in terms of directional or correlational shifts (Tourangeau et al., 2000, p. 198).

For instance, in one of the earliest discovered context effects, the share of respondents who voiced agreement that communist reporters should be allowed to enter and report freely from the United States was significantly higher when the item was preceded by a question that asked whether American reporters should be allowed to enter and report freely from Russia than when the question was asked first (Hyman & Sheatsley, 1950; Schuman & Presser, 1981). This is an example of describing an assimilation effect via a directional shift. Correlational shifts must not necessarily be visible in terms of the mean value of the target item. For instance, Tourangeau et al. (1991) found a higher correlation between reported general and marital happiness when the general question was asked first than when it was asked second.

A closely related but more adaptable distinction is between unconditional, conditional, and associational effects (Rasinski et al., 2012; Tourangeau et al., 2000). An order effect is unconditional when the sheer presence of a preceding question impacts how the respondent thinks about the subsequent one, for instance, by "limiting or directing the way the respondent interprets the subsequent topic" (Rasinski et al., 2012, p. 243). In another example, the sheer presence of prior knowledge questions has been shown to decrease reported topic interest in later questions (Gaskell, Wright, & O'Muircheartaigh, 1995). Conditional order effects imply that the impact of the preceding question on the target question is a function of which response is given to the preceding question. Many classic question order effects in survey research are conditional, including the questions on American and communist reporters and the general-specific questions on general and marital happiness (Smith, 1982, 1992). Self-reported general happiness is higher among respondents who first evaluate their marital happiness—but only among respondents who give positive ratings (Smith, 1992, p. 166). In an associational effect, the strength of the relation between the contextual and the target question is impacted by question order, such as the reported correlational shift between general and marital happiness. Usually, a conditional order effect is associational (Rasinski et al., 2012).

3.2. Explaining context effects in surveys

Beyond merely describing them, researchers have set out to classify context effects according to the underlying steps in the cognitive model of survey response they assume to be impacted (Tourangeau, 1999). When a question concerns an unfamiliar topic or is worded ambiguously, preceding questions are likely to impact the stage of question comprehension. For instance, depending on whether a question about an undefined "educational contribution" was preceded by an item on college tuition or on governmental financial support for students, the term was either understood as a contribution to be paid by students or as financial support given to students, resulting in significantly different levels of support among a student sample (Strack, Schwarz, & Wänke, 1991). In other cases, a context question may impact the information retrieval stage by impacting which information is accessible and whether it is included in the subsequent judgment. In one study, the evaluation of a German political party was significantly more positive when a preceding question framed a highly respected politician as a party member than when the preceding question emphasized the non-partisan office this person held (Schwarz & Bless, 1992). In the study on communist and American reporters (Hyman & Sheatsley, 1950), question order is argued to affect the judgment stage, as neither the interpretation of the question nor the arguments on either side (freedom of the press versus the country's security interests) are impacted by question order. Instead, the judgment prioritizes the respondents' attitude towards communism or the norm of even-handedness (Schuman & Ludwig, 1983).

Moving from a strongly survey-oriented view to a social cognition perspective, the inclusion/exclusion model (Bless & Schwarz, 2010; Schwarz & Bless, 2007) explains context effects based on mental construal. The model's benefits are that it can be used to make predictions about the direction and strength of the effect. Most importantly, however, the model describes when context effects occur. In essence, a judgment requires two mental representations: that of the target or object of evaluation and that of a standard against which the target is evaluated. These mental representations are based on chronically or temporarily accessible information. Contextual information may become part of the mental representation of either the target or the standard. If the information in a previous question is included in the target representation, this results in an assimilation effect, with the answer to the target question becoming more aligned with the answer to the context question. If the contextual information is excluded from the target representation or included in the standard, this leads to a contrast effect. The inclusion/exclusion model contains three filters determining whether contextual information is included in a mental representation. First, the "aboutness" filter decides whether a piece of information comes to the respondent's mind due to an irrelevant influence or whether the piece of information genuinely belongs to their response to the target question. In most cases, contextual information passes the first filter. Next, the representativeness filter checks whether the information is relevant to depict the target. A positive affirmation contributes to the emergence of assimilation effects; if not, the contextual information may be used to construct the standard. The final filter judges the conversational relevance of the contextual information. Respondents assess the common ground (Schober, 1999) that has been gained in the course of the previous question(s), meaning that they consider what they have already communicated, and make a judgment as to whether the researcher wants them to consider the contextual information when responding to the target question, or whether this information would be redundant, thus violating communicative principles (Clark & Haviland, 1977; Grice, 1975). Returning to one of the previous examples, when the question on relationship satisfaction was asked prior to the one on life satisfaction, the correlation between the two questions was much higher when respondents were explicitly requested to include their relationship satisfaction in their evaluation of their overall life satisfaction, than when the lead-in asked them to exclude it (Schwarz, Strack, & Mai, 1991). In self-administered surveys, similar effects have been demonstrated by varying visual design elements, such as presenting the two questions with a box around them or placing them on separate pages (Schwarz, 1996).

The inclusion/exclusion model emphasizes the importance of communicative principles from psycholinguistics in question construal. Most prominently, Grice (1975) described conversation as a cooperative effort and distinguished the four maxims of quantity, quality, relation and manner. The maxim of manner is mainly one of form, requesting that communicators be coherent (in the sense of being brief and unambiguous), and the maxim of quality sets true responses (as opposed to willingly false responses) as a norm. The maxims of relation and quantity are most relevant to the explanation of context effects. The maxim of relation requires that contributions be “relevant” in that they pertain to the preceding contribution. Silber et al. (2020) demonstrated that respondents’ answers to open-ended questions about a preceding survey question directly referenced this question and their survey response in most cases, thus confirming the maxim of relation for a typical probing setting. The maxim of quantity, sometimes referred to as the norm of non-redundancy, requires that contributions be as informative as required, but not more so. Specifically, the underlying given-new contract (Clark & Haviland, 1977) requires that communicators assess which information their dialogue partner already has and not offer redundant information. Applied to the context of surveys, respondents interpret survey questions based on the idea that researchers adhere to these principles when asking questions, meaning that each new survey question asks for additional (and not repetitive) information and, likewise, that giving a “correct” answer means providing the researcher with new information rather than reiterating information already provided in a previous response.

3.3. Context effects in survey settings relevant to web probing

Theories from survey research, social cognition, and psycholinguistics have contributed to explaining the general emergence of context effects in surveys. The following section describes how context effects may emerge in specific survey situations typical of web probing settings. It begins by examining how context effects may emerge in web surveys as a self-administered and computerized survey mode. Next, it introduces response burden as a function of question context. Finally, the differences in cognitive processes when responding to open-ended and closed survey questions and how they differ in potential context effects are discussed.

3.3.1. Survey mode: Web surveys

The mode in which a survey is implemented impacts the context effects that may occur (Doušak, 2017). The effects of the web mode on survey response can be distinguished into aspects of self-administration and computerization (Callegaro et al., 2015, p. 65). The self-administered setting means that respondents may temporarily discontinue filling out a questionnaire, thus interrupting the natural flow of questions (Dillman et al., 2014, p. 327). However, it is aspects of computerization which mainly contribute to web survey-specific context effects, as web survey programming impacts the communicative context of survey questions (Smyth et al., 2007). Relevant design decisions include the grouping of questions, the sequence of questions, and how far respondents can autonomously navigate through the questionnaire.

Web surveys offer all variations of grouping items. For instance, when implementing a multi-item inventory in a web survey, researchers may present each item on a separate screen (known as a paging design), all items on one screen (referred to as a scrolling design), or form several groups consisting of several items each (Peytchev, Couper, McCabe, & Crawford, 2006). Even an entire questionnaire may be presented on one screen, though this approach is uncommon (Dillman et al., 2014, p. 311ff.). Grouping has been shown to affect inter-item correlation, with items presented together on one screen demonstrating higher correlations (Reips, 2002; Tourangeau, Couper, & Conrad, 2004).

Next, the sequence of questions can be randomized to varying degrees. Researchers can install a static question order, in which all respondents receive all survey questions in the same order. A standard procedure is to keep the order of single-item and multi-item measures static, but to randomize items within multi-item batteries. However, potentially all questions within a questionnaire can be randomized. Moreover, researchers may let respondents answer single questions or blocks of questions in the order of their choosing via a menu, resulting in a respondent-selected question order (Callegaro et al., 2015, p. 92; Dillman et al., 2014, p. 314-315).

Closely related to this last option, researchers must decide whether and how respondents may actively navigate between preceding and subsequent questions (Callegaro et al., 2015, p. 90). For instance, respondents may automatically be directed to the following survey page when they have answered the question(s) on the previous page, or they must click a "next" button to get to the next page. Furthermore, researchers may or may not allow respondents to return to previous questions and change their responses. In general, researchers recommend providing respondents with an explicit "next" and "previous" button (Bergstrom, Erdman, & Lakhe, 2016; Couper, Baker, & Mechling, 2011; Dillman et al., 2014, p. 320).

To put it concisely, context effects can potentially flow forward and backward in web surveys so that earlier questions affect later questions and the other way around (an effect also known from paper and pencil and mail surveys, see Schwarz & Hippler, 1995). Whether and how likely such effects are depends on web programming and design decisions regarding question grouping, sequence, and navigation. Therefore, studies using the web survey mode must always consider the implications of web survey programming when examining context effects.

Finally, researchers should be aware that regardless of how they implement their web survey, the presentation can never be entirely consistent across all respondents because "web surveys are now by default multi-device surveys" (de Bruijne, 2015, p. 156). The exact visual presentation is impacted by the device and browser used to access the questionnaire. Respondents may use a large, usually horizontal screen of a PC or laptop or the vertical screen of a mobile device. Researchers, in turn, may choose between a responsive design that presents the optimal visual presentation for a specific device or attempt to keep the visual presentation as consistent as possible across devices (Dillman et al., 2014, p. 307-311).

3.3.2. Response burden and question context

A balance of response (or respondent) burden and motivation is generally considered a prerequisite for high response quality, meaning that a respondent must be willing and able to answer a survey question. The term response burden is often used without an exact definition of what it is, how it can be measured, and which consequences it has for survey data. Importantly, response burden is seldom discussed in relation to context effects, although it strongly depends on contextual factors. This thesis utilizes the definition of response burden by Yan and Williams (2022), in the sense that burden refers to respondents' *perceived* burden, and is impacted by characteristics of the survey and respondent.

Naturally, survey characteristics such as the total survey length or the number of survey pages contribute to response burden (Bradburn, 1978; Groves, Singer, Corning, & Bowers, 1999). When questions consist of many words, include complex linguistic features (Lenzner, Kaczmirek, & Lenzner, 2010), or require extensive retrieval (Yan & Tourangeau, 2008), this increases question difficulty and ultimately, response burden (Yan & Williams, 2022). Relevant respondent characteristics that impact response burden include cognitive ability (Krosnick, 1991, 1999), topic interest (Galesic, 2006), and a generally positive or negative attitude towards surveys (Sharp & Frankel, 1983).

Response burden can be measured by asking respondents about their perceived burden (Galesic, 2006; Yan, Fricker, & Tsai, 2020). More common are indirect measurements of how respondents navigate through surveys by capturing response times via survey paradata (Lenzner et al., 2010; Yan & Tourangeau, 2008) or fixations and pupil dilation via eye-tracking (Neuert, Roßmann, & Silber, 2023; Yan, Williams, Maitland, & Tourangeau, 2016).

Often, response burden is not measured using these direct and indirect measures but rather in terms of its wide range of adverse effects on data quality (Yan & Williams, 2022). These include survey break-off (Galesic, 2006) and item nonresponse (Yan et al., 2020) on the side of survey representation and satisficing behaviour (Hoogendoorn, 2004) on the side of measurement. Satisficing behaviour means giving a satisfactory, but low-quality answer because the respondent does not perform the cognitive tasks required to correctly answer a question (such as retrieving all relevant information from memory). Typical forms of satisficing include giving a “don’t know” answer or choosing the first response alternative that constitutes a reasonable answer (Krosnick, 1991, 1999). One of the heuristics respondents may use to minimize their cognitive effort is to answer a difficult question by substituting it with an easy question (Kahneman, 2012).

In recent years, the perspective that response burden is not static, but dynamic or cumulative, has gained importance, meaning that response burden varies throughout a survey (Read, 2019). The perceived burden may increase throughout a lengthy questionnaire, it may be increased through preceding sensitive or complex questions, but it may also decrease if the topic of a survey section is one the respondent is highly interested in (Yan & Williams, 2022).

To conclude, response burden affects whether and with which motivation respondents answer questions. It is determined by characteristics of the survey, the specific question, and the respondent. Response burden constitutes a “continuous evaluation of the requirements imposed on respondents” (Yan & Williams, 2022, p. 939). Therefore, the survey question and probes themselves, but also the surroundings in which they are asked must be considered to evaluate the burden they may impose. Notably, the response burden associated with probes is generally higher than that of survey questions (Behr, Bandilla, Kaczmirek, Braun, & Majer, 2011). Most prominently, the open-ended format of most probes is described as increasing response burden. The following section discusses the response process of open-ended and closed questions and how this relates to context effects before moving on to the specific cognitive task of probing.

3.3.3. Question format: Open-ended and closed questions

In the most banal terms, the difference between open-ended and closed questions consists solely of the presence or absence of predefined response options. However, the potential implications of question format for survey response are immense. Specific to closed questions is the potential influence of the predefined response options on mental construal and response. Open-ended questions are—ceteris paribus—less specified than closed questions and more prone to the effects of (other) contextual information, such as preceding questions.

There is ample empirical evidence of how response options impact the question-answer process to closed survey questions. For example, response options can be used to infer the pragmatic meaning (Strack et al., 1991) of a question asking how often one is "really annoyed". A low-frequency response scale leads respondents to only consider instances of extreme annoyance, whereas irritations in daily life are included when respondents are shown a high-frequency response scale (Schwarz, Strack, Müller, & Chassein, 1988). Respondents may also use the provided response options to infer typical behaviour, for instance regarding television consumption (Schwarz, Hippler, Deutsch, & Strack, 1985).

Open-ended questions are often associated with increased response burden, particularly regarding the stage of response formatting, when respondents must type in their answer autonomously (Krosnick, 1999; Schuman & Presser, 1979). This reliably leads to any given answer being selected more often in a closed format than in open-ended questions (i.e., Reja, Lozar Manfreda, Hlebec, & Vehovar, 2003). Moreover, other cues provided alongside the question text have a more substantial impact on open-ended than closed questions. For example, Tourangeau, Conrad, Couper, and Ye (2014) demonstrated that respondents asked to name food categories they had eaten were less likely to name these in open-ended than closed questions and that the gap between question formats was smaller for the examples provided in the instructions. This was interpreted as evidence that the instructions had a larger impact on the response in an open-ended format. No empirical evidence specifically examines whether open-ended or closed questions are more prone to the influence of preceding questions.

The primary concern regarding open-ended questions does not lie in differences in response distribution but rather in fears that the cognitive processes involved in construing open-ended questions are influenced by temporarily salient but irrelevant contextual information, with a detrimental impact on the reliability and validity of the responses. Put simply, the worry is that open-ended questions do not measure an underlying opinion as much as random information a respondent happens to be confronted with around the time of the survey. These concerns have been addressed in numerous studies, indicating that these fears are unfounded (Geer, 1991). For example, in experiments that compared responses to open-ended and closed questions on the most important issue facing the country (Schuman, Ludwig, & Krosnick, 1986) or facing the internet (Reja et al., 2003), researchers found that the variety of responses to open-ended questions far exceeded that of closed formats, but that rankings of the responses used in both formats were similar, even when frequency distributions differed markedly. Researchers have argued that differences between responses to open-ended and closed question formats can be minimized when the responses to the closed format are chosen based on question pretests using an open-ended format (Schuman & Presser, 1979).

In current practice, open-ended and closed questions are generally used for different purposes. Closed questions remain the norm, while open-ended questions are implemented when researchers cannot or do not wish to offer predefined response options, for instance, when asking knowledge questions, when they believe predefined response options may bias answers, when they lack knowledge on the range of valid response options or when an exhaustive list would be impractically long (Zuell, 2016). Moreover, researchers argue for using open-ended and closed questions in conjunction to gain deeper insights into the underlying constructs (Silber et al., 2020, cf. Friborg & Rosenvinge, 2013 for a more critical evaluation) and for purposes of question evaluation (Singer & Couper, 2017). This last point brings us directly to the notion of probing.

3.4. The cognitive process of probing

Until now, the focus has been on survey questions. Probes, however, are not merely a type of open-ended survey question or questions on the same topic as a survey question, but questions about the survey question itself. Consequently, the respondents' task lies in thinking about their thinking, also termed metacognition. Metacognition describes "any cognitive process about a different cognitive process" (Overgaard & Sandberg, 2012, p. 1287). It is often referred to as being aware of and regulating one's thinking (Kuhn & Dean, 2010). Introspection can be considered a specific type of metacognition, where the focus lies on conscious experience and is thus subjectively defined (Marcel, 2003). Retrospection is a closely related term emphasizing that respondents must retrieve this information from memory (Fiedler, Ackerman, & Scarampi, 2019). This thesis understands probing as an introspection- and usually retrospection-based task. The first section takes a closer look at intro- and retrospection as question evaluation methods, the fallacies that accompany them, and the implications this has for the survey response process. Based on this, the second part of this chapter contains an evaluation of empirical studies on whether probing is an effective method of examining respondents' cognitive processes.

3.4.1. Using intro- and retrospection as methods of question evaluation

Probing relies on respondent intro- and retrospection to generate data on the cognitive processes underlying their survey response. Probing is originally derived from the method of think-aloud. Think-aloud encourages participants to verbalize their thought processes while they answer a survey question and has been strongly advocated by Ericsson and Simon (1980, 1993) under the term 'verbal protocols'. Think-aloud requires that participants carry out introspection and verbalize their self-observations while simultaneously carrying out the primary task[4] of responding to the survey question. Probing was developed to disentangle these processes by having the respondent carry out the steps consecutively (Converse & Presser, 1986). The respondent first answers a survey question and is subsequently prompted by an interviewer to verbalize the thoughts they had while answering (Willis, 2005). In addition to relying on respondents' ability to introspect, probing requires retrospection, that is retrieving information about the thought processes from short-term memory after carrying out the primary task of answering the survey question.

Undoubtedly, self-reports based on introspection are a highly demanding cognitive task. Therefore, it comes with little surprise that the veracity of self-reports about thought processes has been subject to debate since the onset of experimental psychology, first most prominently in the Wundt-Bühler controversy (Bühler, 1907, 1908; Wundt, 1907, 1908). In a nutshell, Wundt argued that introspection demands the impossible from participants by requiring them to carry out the task at hand and self-observation simultaneously. In contrast, Bühler argued that participants need not actively observe themselves while thinking to retrospectively report on their thought processes, as the information is available in short-term memory (Pongratz, 1997).

[4] It stands to debate whether active self-observation is required during the task or whether a less-demanding inner perception in the sense of Brentano suffices (Brentano, 2014 [1874]; Rodax and Benetka, 2021).

Modern cognitive psychology supports much of Wundt's criticism, distinguishing four main limitations to relying on intro- and retrospection to investigate thought processes (Bröder, 2019; Massen & Bredenkamp, 2005). Survey methodologists have discussed how far these restrictions apply to probing as a method of question evaluation and what needs to be considered when collecting and analysing probe responses (Conrad et al., 1999; Wilson, Lafleur, & Anderson, 1996).

First, some thoughts may be unconscious and simply not detectable through introspection. The methods of intro- and retrospection, and thus probing, are restricted to conscious thought processes that are part of the working memory (Wilson et al., 1996). Instances in which a respondent 'automatically' retrieves information from memory cannot be self-observed. However, researchers agree that most survey responses are consciously constructed rather than automatically retrieved (Tourangeau, 2018, pp. 174-175; Bless & Schwarz, 2010, p. 324) and have argued that answering survey questions mainly involves reportable thinking (Conrad et al., 1999).

Second, introspection may lead to reactivity[5], meaning that self-observation may alter the thinking process. Reactivity has been demonstrated for problem-solving tasks for both think-aloud (Russo, Johnson, & Stephens, 1989) and retrospective verbalization (Schooler, Ohlsson, & Brooks, 1993). Applied to the task of responding to survey questions, reactivity through introspection implies that respondents process and answer survey questions differently when probes are embedded than when they are not. One way this may occur is that the depth of processing of survey questions is impacted. Respondents generally give "fast" and spontaneous answers to survey questions; in fact, instructions often require respondents to answer spontaneously. Probes, by contrast, promote a "slow" route (Kahneman, 2012) to responding to survey questions, described as System 2 in terms of dual-process theories (Evans, 2003), or as the systematic (Chen & Chaiken, 1999), or central route (Petty & Cacioppo, 1986). The exact implications of reactivity for survey data have yet to be discussed; however, Wilson et al. (1996) report a study in which test-retest reliability for an attitude question was lower in a group of respondents who were asked to verbalize their thoughts on the subject prior to answering

[5] By nature, survey questions already constitute a reactive measurement, as demonstrated by Knowles (1988) and Knowles et al. (1992). In this sense, probing adds an additional layer of reactivity to an already reactive method.

the survey question than in a group of respondents who were only presented the survey question.

Third, retrospection is prone to memory errors[6], limiting the accuracy of self-reports. This means that participants might fail to report thought processes that occurred, or report thought processes they did not have. Empirical evidence of incorrect recollection of thought processes is available for multiple psychological research areas. Bredenkamp (1990) demonstrated that a mathematical wizard made wrong assumptions about the cognitive steps he applied to solve complex mathematical equations and which calculation steps were particularly demanding. Nisbett and Wilson (1977) reviewed empirical evidence on consumer decisions and attitude change, concluding that self-reports seldom reflected the individuals' thoughts or problem-solving steps. Instead, self-reports on thought processes are based on implicit or general theories about how one might arrive at a particular conclusion or decision. Some researchers have argued that the directive method of probing fosters such post-rationalization more than think-aloud (Fox, Ericsson, & Best, 2011). Wilson et al. (1996) conclude that probing is unlikely to uncover reasons not part of a respondent's causal theory when explaining one's choice of survey response.

Finally, intro- and retrospection require the task of verbalizing self-observed thoughts. Thoughts may not be saved in words but as images and must first be transformed into language (Bröder, 2019). In web probing, they must additionally be transferred into writing (Behr et al., 2017). The burden of these steps increases the probability that respondents only disclose thoughts that are easy for them to verbalize (Wilson et al., 1996).

In light of these limitations, modern cognitive psychology advocates for using intro- and retrospection to *generate* hypotheses about thought processes, but these hypotheses should be confirmed using other methods, such as outcome-based methods or other process-oriented methods using objective measures (Bröder, 2019). Several studies have confirmed the benefits of evaluating survey questions by combining introspection and other question evaluation methods (Maitland & Presser, 2016, 2018; Presser et al.,

[6] Again, survey responses themselves are likely to include memory errors, caused be challenges in encoding, storing, retrieving, and reconstructing events (see Tourangeau, 2000, for a detailed description of memory errors in surveys).

2004). Despite this, survey researchers argue that the benefits of collecting the respondents' view outweigh the risks of collecting partially inaccurate data (Beatty & Willis, 2007).

3.4.2. The efficacy of probing as a method of question evaluation

In light of the fallacies involved in intro- and retrospection, assessing the efficacy of probing as a method of question evaluation seems necessary. To this end, empirical evidence is required to establish whether probing achieves its goals. When probing employs a descriptive approach, data collected via probes should depict how respondents construct the pragmatic meaning of a survey question (Miller, 2014). When probes are implemented following a reparative approach, their primary goal is to successfully identify problems respondents encounter during survey response (rather than artifacts from the probing context). Furthermore, question revisions that rectify these problems should be derived from the findings (Willis, 2015a).

Regarding the descriptive approach, it remains challenging to assess whether probe responses depict respondents' mental construal of a survey question at the time of survey response. In recent years, several studies have used coded web probe responses to explain the responses to the closed survey items. The rationale of these studies is that if respondents can verbalize (and type) their thought processes in answer to web probes, the coded probe responses should show a strong relation to the survey responses. Behr, Braun et al. (2014) demonstrated that web probing results could explain the variance of an item on civil disobedience. Using regression analysis with the response to the survey item as a dependent and the substantive codes from web probing as independent variables, they established that general dissatisfaction with the government and a perceived lack of politicians' responsiveness towards voters' needs were strongly associated with support of civil disobedience. In contrast, an understanding of civil disobedience as including violence and destruction was associated with disagreement with the item. In another study, probes were used to collect reasons for seemingly contradictory responses to two survey items on traditional and egalitarian gender ideology. The reasons named by respondents aligned well with different response patterns to the survey questions (Behr, Braun et al., 2012). While these findings are encouraging, the limitation of these studies is that post-rationalization of the survey responses (i.e., Bröder, 2019; Nisbett & Wilson, 1977) would likewise explain the results.

In light of the difficulties in verifying the descriptive approach to probing, it comes as no surprise that the vast majority of studies on the efficacy of probing have focused on the method’s reparative goals (see Willis, DeMaio, & Harris-Kojetin, 1999, for an overview). To assess whether the problems identified by probing depict actual problems with a survey question, researchers have examined whether the problems translated into overt survey response behaviour and how the problems found by probing compare to findings from other question evaluation methods. To assess whether findings from probing contribute to improving survey data quality, split-ballot experiments have compared survey questions revised after probing with survey questions prior to revision.

Findings from probing have proven effective at identifying problems that translate into survey response behaviour in several empirical studies. Forsyth, Rothgeb, and Willis (2004) found that problems detected by various pretesting methods, including probe-based cognitive interviewing, predicted item nonresponse, and overt behaviour, such as asking for clarifications. Using web probing data, Meitinger (2018) found problem codes, such as issues of comprehension or sensitivity, to be strong predictors of survey response for a single-item measure of general national pride.

Studies that have compared the problems identified by probing with those identified by other question evaluation methods have found mixed results. Several studies compared probing in cognitive interviews with other methods, such as behaviour coding, expert reviews, or latent class analysis (Presser & Blair, 1994; Willis, Schechter, & Whitaker, 1999; Yan et al., 2012). However, methods of question evaluation “differ in their underlying assumptions, the data collection methods they use, [and] the types of problems they identify” (Yan et al., 2012, pp. 503-504). Therefore, it comes as little surprise that studies have generally found, at best, moderate consistency of results.

Following the development of web probing, several studies have compared web probing results to findings from probe-based cognitive interviews (Behr, Braun et al., 2012; Lenzner & Neuert, 2017; Meitinger & Behr, 2016). The argumentation behind this approach is that findings from cognitive interviews have been shown to correspond to problems in production surveys (Lenzner, Hadler, & Neuert, 2022; Willis & Schechter, 1997); therefore, if findings from web probing correspond to those from cognitive interviews, this speaks in favour of web probing being an effective method of question evaluation. For instance, in a study by Behr, Braun et al. (2012), web probes replicated findings from probe-based telephone interviews reported by Braun (2008). In another study, Meitinger and Behr (2016) evaluated a ten-item battery on specific national pride using either web probing or cognitive interviews. Despite differences in the level of probe nonresponse and the length of responses, they found that the methods had “an extensive overlap of results” (Meitinger & Behr, 2016, p. 376) regarding the themes named and problems found with the items. Similarly, Lenzner and Neuert (2017) found that web probing and cognitive interviews detected similar themes and errors for measures of national identity and citizenship, leading to similar recommendations regarding question revision.

Naturally, when probing identifies problems with survey questions, these findings are used to create question revisions to improve data quality. Therefore, several studies have conducted split-ballot experiments with original and revised survey questions. One study compared the correlation of original and revised measurement instruments with related constructs using revisions based on cognitive interviews (Lenzner et al., 2022), while another study compared the reliability and cross-cultural measurement invariance of instruments prior to and after revisions based on either cognitive interviews, web probing or expert reviews (Menold, Hadler, & Neuert, 2023). Unfortunately, both studies yielded ambivalent results regarding the efficacy of probing to generate higher-quality survey data. The limitation of this approach is that it rests on the assumption that probing not only identifies problems with survey questions but also points to effective remedies to these problems.

To sum up, evidence supports the notion that probe responses from both cognitive interviews and web probing can be used to predict problems with survey questions that translate to overt survey response behaviour. However, there is likewise evidence that the method of intro- and retrospection may lead to fallacies on the level of individual respondents, and the results of studies employing web probing to predict response distributions can alternatively be explained with the effects of post-rationalization. Therefore, pleas to combine cognitive pretesting methods with other forms of question evaluation that rely on objective measures (Behr et al., 2020; Benítez & Padilla, 2014; Benítez, van de Vijver, & Padilla, 2022; Benítez Baena & Padilla, 2014) seem more than reasonable. At the least, probing delivers valuable insights to generate hypotheses about respondents’ thought processes while answering survey questions. At the same time, it becomes clear that respondents do not have direct and error-free access to their mental processes, which they can translate into (written) probe responses. Instead, survey and probing questions form a communicative context in which a cooperative respondent is inclined to give consistent answers. The following section describes how this context impacts survey and probe responses.

4. CONTEXT EFFECTS IN WEB PROBING

The following chapter applies the current state of knowledge about the mechanisms underlying survey context effects to settings involving web probing. It begins by classifying context effects in web probing in terms of the potential directions of effects and the possibilities of combining survey and probing questions. Following this, the second subchapter introduces a psychological model of context effects in web probing. The mechanisms underlying the model's three directions of effects are described based on the cognitive and communicative principles from Chapter 3; moreover, previous empirical research in these areas is discussed and existing research gaps are identified. The final section derives three research questions from these gaps and closes with an overview of the empirical studies.

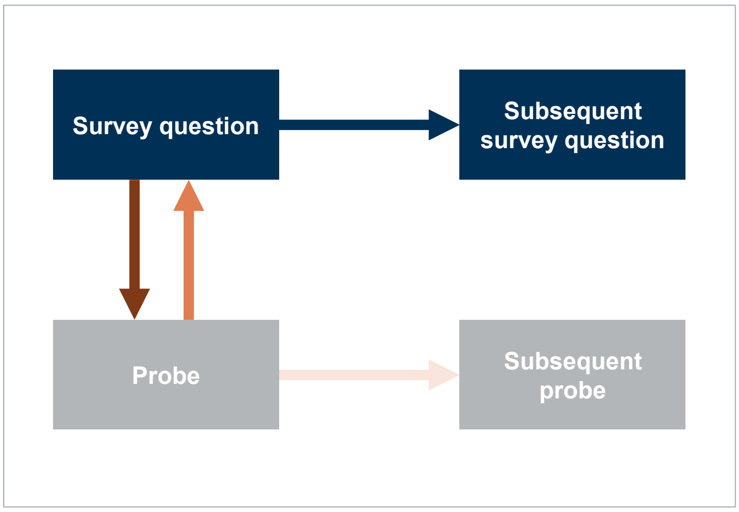
4.1. Describing context effects in web probing

Context effects examine how the contextual setting impacts the response to a target question. In research on survey questions, this often means examining whether and how a target question is affected by the presence or absence of a preceding question. Examining context effects in web probing produces research designs that are more diverse and complex.

4.1.1. The direction of context effects

The most fundamental distinction between research on context effects in web probing and research that focuses on survey questions is that there is no single answer as to which question (or which questions) constitute the target question(s). Leaving out the effects of survey questions on other survey questions—which is the focus of context effects in surveys—three general directions of effects can be distinguished (see Figure 4.1).

The first direction of effects is that of survey questions on probes, as indicated by the dark orange arrow. These effects include the general impact of the survey question on the probe pertaining to it, the specific effect of the chosen survey response, and the effects of intermittent survey questions, that is survey questions placed between a survey question and the related probe in a retrospective probe design.



Note. Orange arrows indicate context effects in web probing; the dark-blue arrow indicates context effects between survey questions.

Figure 4.1. Context effects in survey and web probing research

The second direction of effects is the impact of probes on survey questions, indicated by the medium orange arrow. This includes the effects of probes on response behaviour to the survey question the probe pertains to, but also on subsequent survey questions. If probes impact survey response behaviour, this may be visible in the responses to single survey items or the strength of the relation between two or more items.

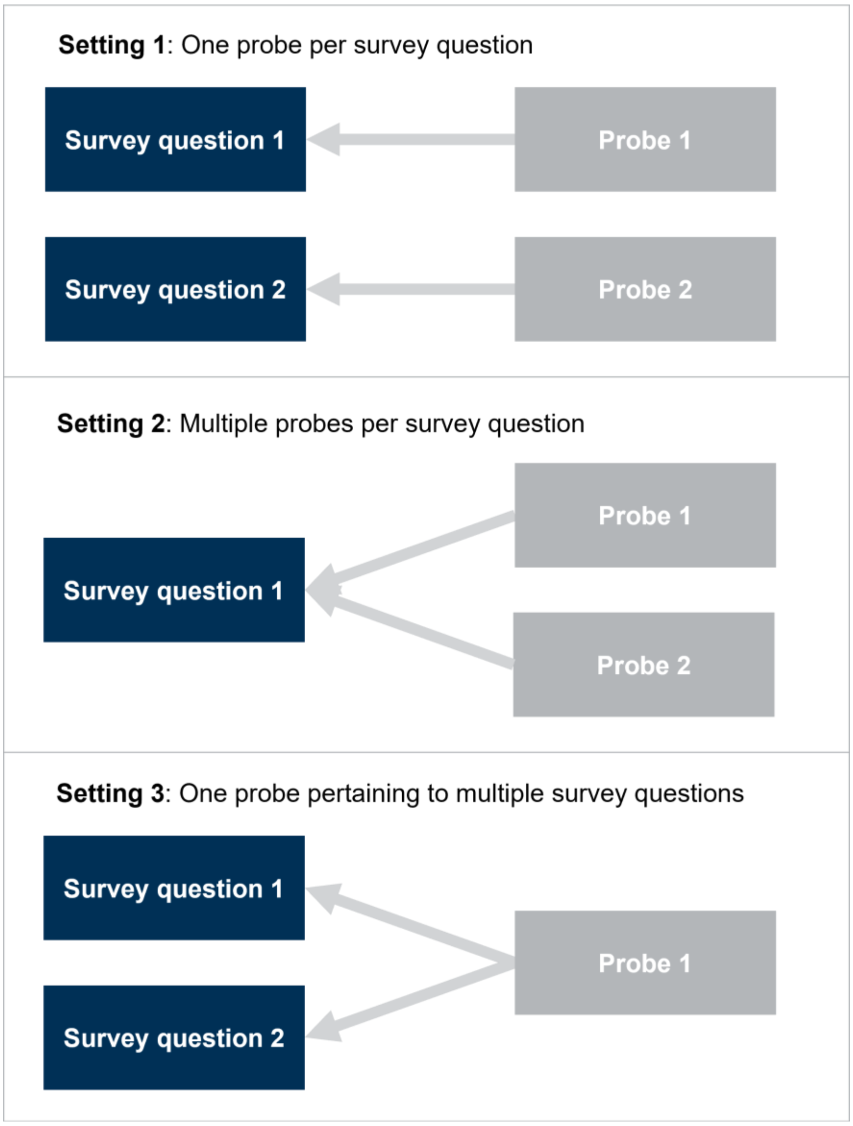
The third direction of effects is that of probes on each other, indicated by the light orange arrow. To name just a few potential manifestations, respondents may try to give consistent probe responses or may hesitate to offer probe response content that they consider redundant. Notably, the different directions of effects do not exclude each other, but may coincide.

4.1.2. Research designs in web probing

Research on survey context effects usually focuses on two survey questions: the target question and the context question, such as in classic experiments on life and relationship satisfaction (Schuman & Presser, 1981; Smith, 1982). The only common deviation from this setting is when the target question is accompanied by several contextual questions, for instance, when work and leisure satisfaction are included as additional specific domains (Schwarz, Strack, & Mai, 1991). Research on context effects in web probing must consider that the number and combination of questions can vary. Moreover, all questions can potentially be influenced by surrounding questions.

Question constellation: Combinations of survey and probing questions

Research settings on context effects in web probing must include two question types: survey questions and probes. One can distinguish between three possible question combinations in web probing (see Figure 4.2).


Setting 1: One probe per survey question
Survey question 1
Probe 1
Survey question 2
Probe 2
Setting 2: Multiple probes per survey question
Probe 1
Survey question 1
Probe 2
Setting 3: One probe pertaining to multiple survey questions
Survey question 1
Probe 1
Survey question 2

Note. Arrows indicate which survey question(s) the probe(s) pertains to

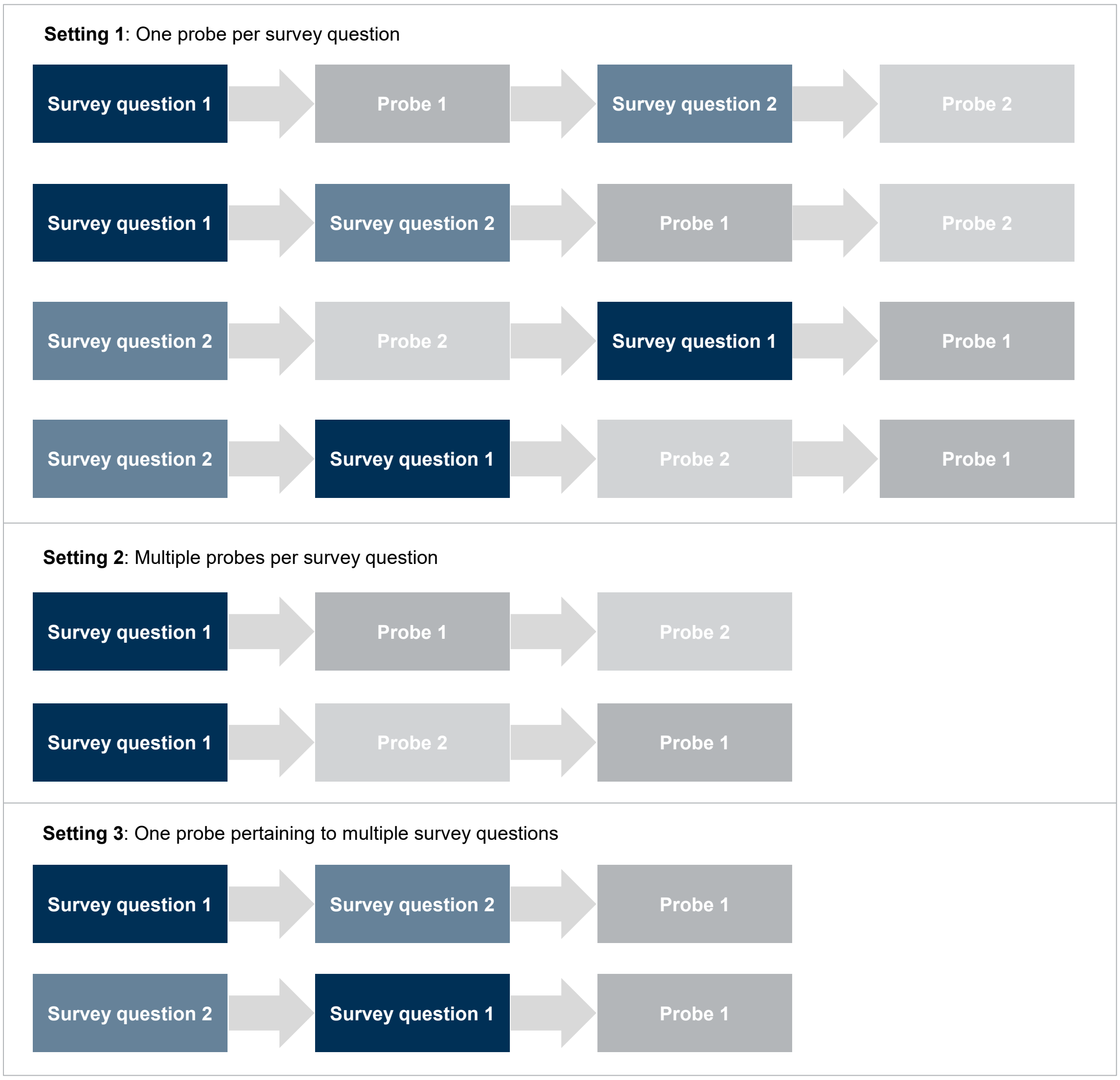
Figure 4.2. Question constellation in web probing

The first possibility is to employ one probe per survey question. At its simplest, this means having one survey question and one probe about it, or as shown in the graphic, several survey questions, each accompanied by one related probe. The second possible combination of survey questions and probes is to have multiple probes that pertain to the same survey question. For instance, a survey question may be followed by a category selection and a specific probe (i.e., Meitinger et al., 2018). The third setting is to have multiple survey questions followed by one probe pertaining to all survey questions. An example of this setting is when probes are used to examine contradictory response patterns across survey questions (i.e., Behr, Braun et al., 2012) or when a probe examines a term used in multiple preceding survey questions (Braun et al., 2013).

Question order: Sequence of survey and probing questions

Several question sequences are conceivable for each of the three constellations described above, depicted in Figure 4.3. Two important aspects of question sequence must be considered for constellations involving one probe per survey question. The first is the aspect of probe placement (see Chapter 2.2). Suppose concurrent or embedded probing is used, and probes directly follow the survey questions they pertain to (as depicted in the first and third example in setting 1). Here, the flow of the survey questions is interrupted by intermittent probes. In these cases, the main research interest lies in examining the *effects of probes on survey questions*. Suppose retrospective probing is employed (second and fourth example). Here, the natural flow of the survey questions is maintained but there are intermittent survey questions between a probe and the survey question it pertains to. In these cases, the *effect of survey questions on probes* will be the primary focus of the analysis.

However, the first setting also includes research designs in which the order of the survey questions is varied. In the upper two cases, survey question 1 and probe 1 are asked before survey question 2 and probe 2; in the lower two cases, the order is reversed. This change in the order of survey questions mirrors the typical randomization carried out in research on context effects for survey questions. Notably, the change in the order of the survey questions is mirrored by a change in the order of the probes. In these cases, the main research focus will be the *effect of probes on each other*, as the first-shown probe might affect the response given to the second-shown probe.



Note. Arrows indicate which question directly follows the previous one

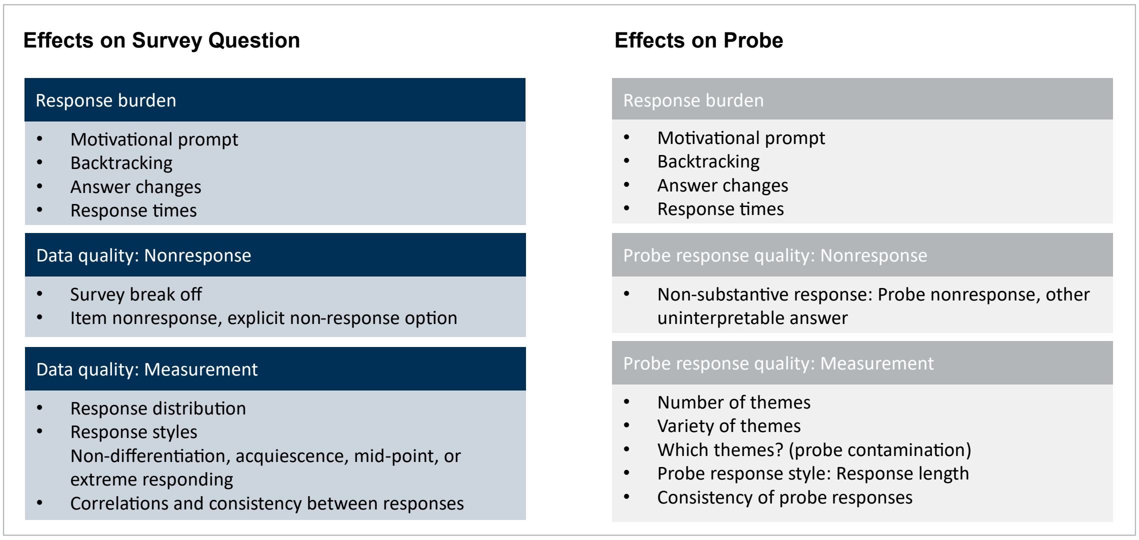
Figure 4.3. Question order in web probing

For the second setting, which employs multiple probes about one survey question, these probes' sequence may vary. In this case, the main research interest will also lie in the effects of probes on each other.

Finally, the order of the survey questions may vary for settings involving one probe pertaining to multiple survey questions. This case will not be discussed further, as it is not so much a matter of context effects in web probing but of examining context effects of survey questions using web probing (see Bishop, 1992; Bishop, Oldendick, & Tuchfarber, 1985 for studies using think-aloud to examine question order effects).

4.1.3. Manifestations of context effects in web probing

Context effects may manifest in survey and probe responses in many ways. Previous research has mainly focused on survey response behaviour, such as response distributions, or measures of the relation between survey responses, such as correlations (see Chapter 3.1). Further parameters may enhance our understanding of the mechanisms underlying context effects in survey research and web probing alike. For instance, research by Knowles et al. (1992) demonstrated that question order effects in personality measures impacted response times to questions and the tendency to extreme responding. Moreover, the existing parameters require adjustment and expansion when examining the responses to open-ended questions, the relation of open-ended responses to one another, or the relation between responses to open-ended and closed questions. Although it is beyond the scope of this thesis to examine or even enumerate all potential measures, the following taxonomy may contribute to systematically examining context effects in web probing (see Figure 4.4). For both survey questions and probes, it distinguishes between indirect measures of (perceived) response burden (Yan & Williams, 2022; see Chapter 3.3.2) and overt effects on data quality in the sense of nonresponse and measurement error (Groves et al., 2011).



Effects on Survey Question

Response burden

- Motivational prompt
- Backtracking
- Answer changes
- Response times

Data quality: Nonresponse

- Survey break off
- Item nonresponse, explicit non-response option

Data quality: Measurement

- Response distribution
- Response styles
  Non-differentiation, acquiescence, mid-point, or extreme responding
- Correlations and consistency between responses

Effects on Probe

Response burden

- Motivational prompt
- Backtracking
- Answer changes
- Response times

Probe response quality: Nonresponse

- Non-substantive response: Probe nonresponse, other uninterpretable answer

Probe response quality: Measurement

- Number of themes
- Variety of themes
- Which themes? (probe contamination)
- Probe response style: Response length
- Consistency of probe responses

Figure 4.4. Manifestations of context effects in web probing

Regarding response burden, survey navigation paradata (Heerwegh, 2011; Kreuter, 2013) can measure how respondents react to the target survey question or probe.

Depending on the research design and question, measures may include the activation of motivational statements to determine whether respondents try to leave a survey question or probe unanswered (Kaczmirek, Meitinger, & Behr, 2017), data on backtracking and answer changes (Heerwegh, 2003), or response times (Yan & Tourangeau, 2008).[7]

Second, any contextual factor in web probing may impact representation by increasing or decreasing nonresponse error (Groves et al., 2011; Groves & Lyberg, 2010) in the form of survey break-off or item/probe nonresponse. Past research on question order effects has not examined the effects on survey break-off. However, at the latest when open-ended probes with high response burden are part of the research design, survey break-off becomes a relevant aspect of survey data quality (Galesic, 2006; Peytchev, 2009; Roßmann, Blumenstiel, & Steinbrecher, 2015). Contextual factors in web probing may impact nonresponse to survey questions in the form of refusals or “don’t know” answers (Shoemaker, Eichholz, & Skewes, 2002; Yan & Curtin, 2010). Regarding probe responses, respondents may choose to leave a probe unanswered or only offer non-substantive content, such as “don’t know” answers and off-topic remarks (see Behr et al., 2012b for a detailed differentiation of non-substantive probe response content).

Moreover, the quality of the measurement may be impacted by contextual factors. Regarding survey data, response distributions to single survey questions are of interest, as well as response behaviour such as straightlining (Kim, Dykema, Stevenson, Black, & Moberg, 2019) or response styles such as acquiescent, disacquiescent, mid-point or extreme responding (van Vaerenbergh & Thomas, 2013). For multi-item measures, sum scores, correlations between responses to two or more survey questions, factor loadings, and further measures of inter-item consistency (Huang, Curran, Keeney, Poposki, & DeShon, 2012) may differ as a function of context. Probe response content can be examined via the total number and the variety of themes. Memory errors may be examined via the likelihood of naming specific themes depending on the preceding question(s). Probe responses may vary in response styles, such as response length or theme density (Kunz & Meitinger, 2022). Finally, responses to several probes may be consistent or contradictory to each other (i.e., Lee et al., 2020).

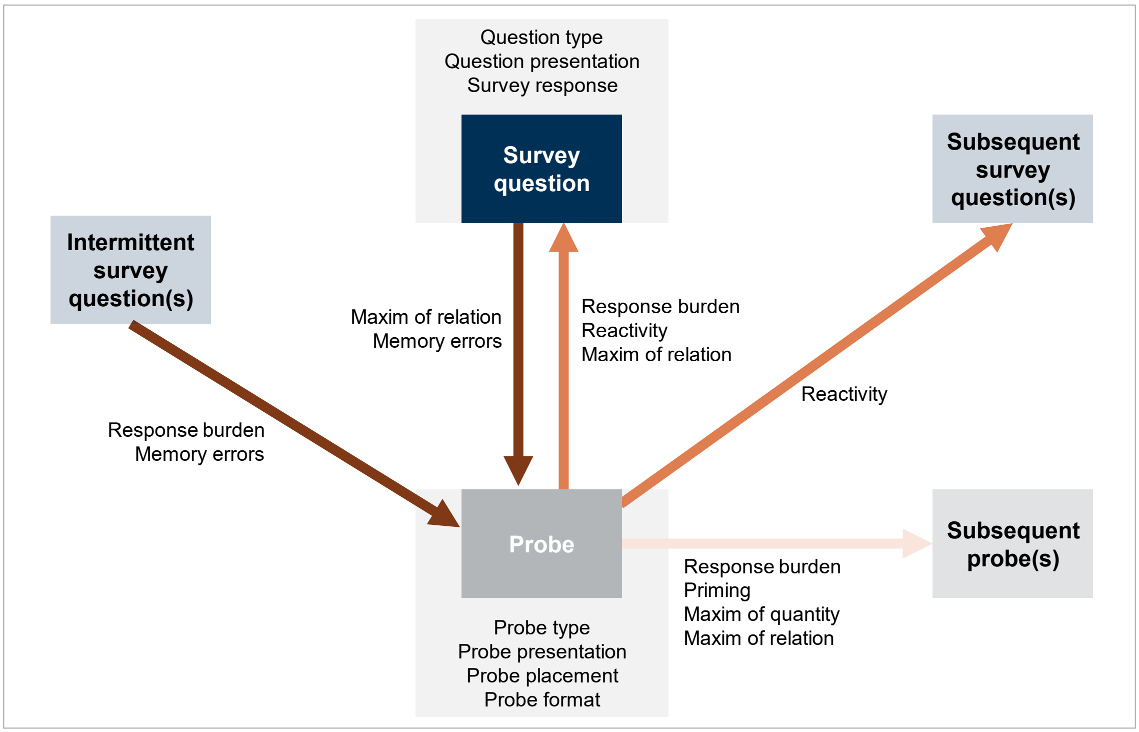
[7] Eye-tracking could be added to this enumeration as a typical indirect measure of response burden (Neuert et al., 2023; Yan et al., 2016). It has been left out as it is currently not possible to collect eye-tracking data without disrupting the self-administered web probing setting.

4.2. Towards a model of explaining context effects in web probing

Based on the cognitive and communicative principles that guide context effects in surveys and the cognitive processes specific to probing, the assumed mechanisms underlying context effects in web probing are the communicative *maxims of relation* and *quantity*, *response burden*, *reactivity*, and *memory errors*. The other fallacies of introspection-based question evaluation—the inability to retrieve unconscious thought processes and challenges verbalizing complex thoughts—are considered general limitations of cognitive probing but are assumed to be mainly independent of question context. Guided by *communicative principles,* respondents try to give coherent, and thus relevant, answers to survey questions and probes (applying the *maxim of relation*) and make inferences about which content can be considered new information to the researcher (applying the *maxim of quantity*). Respondents' motivation to be cooperative communicators is moderated by the perceived *response burden* a given web probing design inflicts upon them. Probes require that respondents engage in introspection. This results in *reactivity*, meaning that how respondents mentally process survey questions is impacted when they are (or expect to be) asked to reflect their responses. Because respondents do not always know or cannot always remember why they responded in a certain way, *memory errors* caused by the retrospective task inherent to most probing designs manifest themselves.

Potential types and directions of context effects in web probing are illustrated in Figure 4.5**.** The dark blue survey question and dark grey probe present the minimum number of questions necessary to examine context effects in a setting employing probing. Intermittent and subsequent survey questions are indicated in light blue, and subsequent probes in light grey. The arrows indicate possible context effects and are labelled with the underlying mechanism(s) of these effects. The colour of an arrow indicates the direction of an effect.

In a nutshell, the survey question—including all its defining characteristics, such as question type, presentation, and survey response—sets the frame for a probe pertaining to it. The probe type, presentation, placement, and format determine the precise cognitive task required of the respondent. Respondents usually function as cooperative communicators, trying to provide a relevant answer to the probe. Suppose they cannot access the information required from them (*memory errors*). In that case, the model postulates that they are likely to create an answer coherent with the given survey response (*maxim of relation*), resulting in post-rationalization. When probes are not presented alongside the question they pertain to or directly following it, the *response burden* caused by the retrospective probing task increases and respondents' access to their short-term memory decreases (*memory errors*). Therefore, they become more likely to draw on contextual information, for instance, if intermittent survey questions offer relevant content to draw on.



Note. Dark orange arrows indicate the effects of survey questions on probes; medium orange arrows indicate the effects of probes on survey questions; the light orange arrow indicates the effects of probes on each other.

Figure 4.5. Model of context effects in web probing

Taking on the opposite direction of effects, a probe implemented into a web survey may impact the question it relates to. Adding a probe to a survey question increases *response burden*. If the survey question and probe are presented in an embedded design, respondents may react differently to the survey question due to the probe (*reactivity*). However, even when respondents process the survey question and probe consecutively, the probe may cause them to return to the survey question and align their survey response to match the thoughts they had during probing (*maxim of relation*). Moreover, respondents who have already answered one or more probes may react differently to subsequent survey questions as they expect to be interrogated about them (*reactivity*).

Finally, probe responses are not independent of each other. As open-ended probes pose a burdensome respondent task, the *response burden* increases with each probe. Respondents learn to expect a particular type of probe when a web survey includes multiple probes of the same type. They will likely offer probe response content matching the previous probe types, even when a different task is required (*priming*). Moreover, respondents may take care not to contradict themselves in different probe responses (*maxim of relation*) while simultaneously trying to offer new information in answer to each probe (*maxim of quantity*). The following section describes these effects and how far previous research has examined them.

4.2.1. The impact of survey questions on probes

The first perspective to be taken on is that survey questions impact probes. Generally, one can distinguish between (1) the effects of a survey question on the probe pertaining to it and (2) the effects of intermittent survey questions in retrospective probing designs.

The impact of a survey question on a probe that pertains to it

That a survey question impacts a related probe may appear banal but constitutes the foundation for probing: the question under examination sets the frame for the corresponding probe. This applies regardless of whether a probe is embedded alongside the survey question on the same page, concurrently on the following survey page or retrospectively (see Figure 4.6).

This effect can function as an *unconditional effect* (Rasinski et al., 2012; see Chapter 3.1) caused by the *maxim of relation* (Grice, 1975; see Chapter 3.2). A probe response must relate to the preceding survey question to be relevant. Supporting this, there is empirical evidence that the vast majority of open-ended probe responses directly refers to the content of the survey question (Friborg & Rosenvinge, 2013; Silber et al., 2020). Respondents demonstrate a need to explain or even justify their chosen survey response when answering probes (Meitinger et al., 2022). This stable mechanism is the basis for analyses that employ coded probe responses to explain survey response behaviour (i.e., Behr, Braun et al., 2014; Meitinger, 2018).

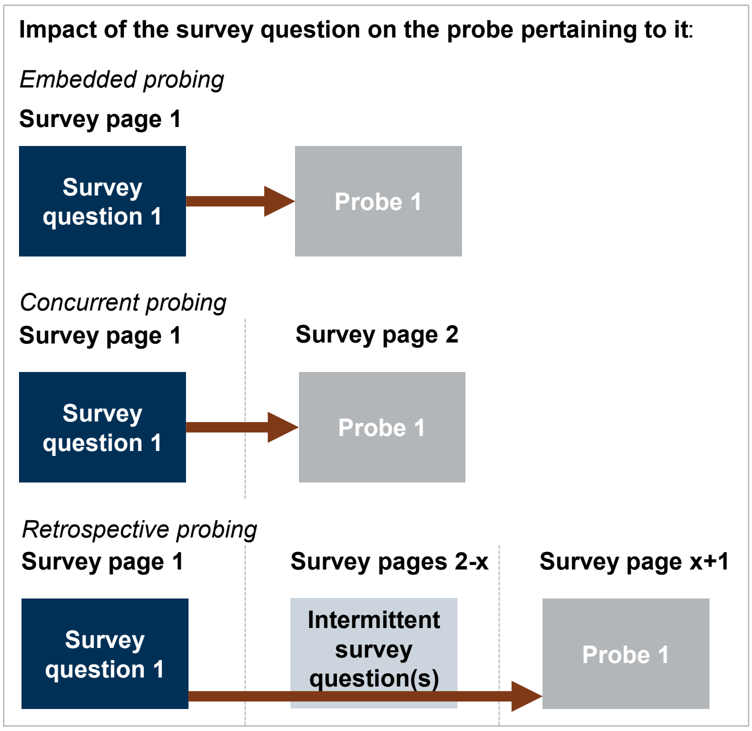


Figure 4.6. Impact of the survey question on the probe pertaining to it

In addition, there is indication that the likelihood of providing a high-quality probe response varies by the chosen survey response. In an interviewer-administered study on left-right orientation, Zuell and Scholz (2015) demonstrated that respondents' self-reported political orientation predicted the likelihood of giving substantive responses to probes on the understanding of the terms "left" and "right". Respondents who placed themselves on the "left" side of the political orientation scale (rather than on the right side or in the middle) were more likely to give substantive responses to comprehension probes for both terms. This constitutes a *conditional effect* of the survey question on the probes. There are several possible explanations for this effect. For one, the chosen survey response may be indicative of respondent characteristics that promote high-quality probe responses, such as topic interest (Holland & Christian, 2009), attitude strength and accessibility (Krosnick & Smith, 1994, p. 280), a generally positive or negative attitude towards surveys (Rogelberg, Fisher, Maynard, Hakel, & Horvath, 2001), personality traits such as conscientiousness (Zuell & Scholz, 2015), or education (Schmidt, Gummer, & Roßmann, 2020). Alternative explanations that directly pertain to the survey response are likewise possible. For instance, some survey responses may inherently require more explanation than others for self-presentation (see Couper, 2013; Singer & Couper, 2017 for similar argumentation). However, the prevalence and underlying mechanisms of such conditional effects of survey questions on probes have yet to be empirically examined.

The effects of the survey question and the chosen survey response on the respective probe are likely to play a crucial role when *memory errors* occur (Bröder, 2019; Massen & Bredenkamp, 2005; see Chapter 3.4), for instance, if respondents do not know or remember why they chose a particular survey response (Wilson et al., 1996). This can happen when a survey response is generated ‘automatically’, and the process is therefore not part of the respondent’s working memory. Especially in retrospective probing, the reasons for responding to the survey question may no longer be retrievable. The *maxim of relation* assumes that respondents will give a relevant answer despite this, meaning that they provide a probe response consistent with the chosen survey response through post-rationalization.

In summary, probe responses generally pertain to the survey question under investigation and are consistent with the chosen survey response (*maxim of relation*). In how far this is because respondents can access and report on their cognitive processes, or whether respondents encounter *memory errors* and post-rationalize is challenging to assess. Research on probe response quality has indicated that the likelihood of giving an interpretable probe response may depend on the selected survey response (Zuell & Scholz, 2015). It has not been examined whether this is a common effect or what causes it.

The impact of intermittent survey questions on a probe

The second line of conceivable effects of survey questions on probes are the effects of survey questions *other* than those under investigation. The prerequisite for such effects is that probes are placed retrospectively, meaning that they are asked later in or at the end of a survey rather than on the same survey page as the question they pertain to (embedded placement) or on the very next survey page (concurrent placement) (see Figure 4.6). Retrospective probing is common practice in both cognitive interviewing (i.e., Willis & Artino, 2013; Willson, Cibelli Hibben, & Gregory-Lee, 2022; Willson & Miller, 2022) and web probing (Fowler et al., 2016; Fowler & Willis, 2020) and is recommended to maintain the natural flow of the survey questions (Collins, 2015) and to prevent effects of probes on subsequent survey questions (Drennan, 2003; Willis, 2005).

From a theoretical perspective, intermittent survey questions increase perceived response burden and memory errors. The increased *response burden* is due to the disrupted natural flow of conversation. In retrospective probing, respondents have moved on to other survey questions—potentially to completely different topics—and are then asked to return to a previous topic. As a result, they must closely re-read the survey question text and invest additional effort to carry out retrospection. The increased response burden will likely result in longer response latencies and a higher likelihood of respondents leaving probes unanswered.

Memory errors caused by intermittent survey questions can occur as an unconditional or conditional effect. The *unconditional effect* is that the distraction caused by intermittent survey questions results in respondents having less content in their short-term memory. Less available content to report should result in increased non-substantive probe responses, fewer mentioned themes and less variety in the themes named. A web probing study reported by Fowler and Willis (2020) compared the probe response content of probes asked directly after an item battery with ones asked retrospectively at the end of a survey. They found a significantly higher share of relevant responses to one of four probes in the concurrent condition, and a slight trend towards longer responses in the retrospective condition. However, the difference was only significant for the first probe. It must be noted that their study was not a randomized experiment as the conditions were fielded several weeks apart. In cognitive interviewing, studies have indicated that think-aloud and concurrent probing lead to higher-quality responses than retrospective probing. One study found that think-aloud and concurrent probing detected a similar number of problems, while retrospective probing uncovered markedly fewer problems (Daugherty, Harris-Kojetin, Squire, & Jaël, 2001). In product decision-making, think-aloud interviews generated more relevant information, particularly insights into cognitive steps and difficulties encountered during decision-making. At the same time, retrospective probes delivered more insights into the final decision (Kuusela & Paul, 2000). Similarly, think-aloud produced more procedural information in a usability study, whereas retrospective probing produced more explanations for the final behaviour (Bowers & Snyder, 1990).

The second potential effect of *memory errors* caused by intermittent survey question(s) is *conditional*, as its occurrence depends on the topical relation of the intermittent survey question(s) with the survey question under investigation and the retrospective probe. In this case, respondents may use the related intermittent survey question(s) as memory cues. For instance, in a general-specific question combination (i.e., on life and relationship satisfaction), respondents first answering a general and then a specific question may falsely remember the topic of the specific question as a relevant aspect of the general question when answering a retrospective probe, even if this topic was in truth not part of the respondent's mental processes at the time of answering the general question. Cognitive pretesting practitioners have long been aware of the possibility of such effects (Collins, 2015, p. 120), though there have been no respective empirical studies. Despite this, retrospective placement is mentioned in nearly every introduction to cognitive pretesting (i.e., Snijkers, 2002, p. 77-84) as a way to "elicit valid information about the thought processes used by respondents" (Drennan, 2003, p. 60). Active recommendations to employ retrospective probes are often restricted to situations which prioritize not interfering with the flow of the questionnaire, such as when testing self-administered questionnaires (Willis, 2005, p. 51f.), or when concurrent probing may impact the response to subsequent, related questions (Collins, 2015, p. 120). The last case implies that when several survey questions relate to each other, retrospective placement should be preferred.

In summary, previous research has indicated an unconditional effect of intermittent questions on memory errors in later probe responses. No studies have examined whether respondents in retrospective probing suffer from increased perceived response burden in the form of longer response times or whether there is an increase in attempts to leave probes unanswered. Notably, no research to date has examined the possible conditional effects of intermittent survey questions. If such effects occur, intermittent survey questions would serve as memory cues when retrospective probes are asked at the end of a survey section containing topically related questions.

4.2.2. The impact of probes on survey questions

Possibly the first question researchers considering employing web probes ask themselves is whether implementing probes will impact their survey. Regarding this direction of effects, one can distinguish between (1) the effects of the probe on the survey question it pertains to and (2) the effects on other, subsequent survey questions.

The impact of the probe on the survey question it pertains to

A probe may be presented on the same survey page as the question it relates to, referred to as embedded probing, or on a subsequent page in a paging design, typically in the form of concurrent probing. Depending on how a probe is presented, the effects of response burden, reactivity, memory errors and the maxim of relation may apply differently (see Figure 4.7).

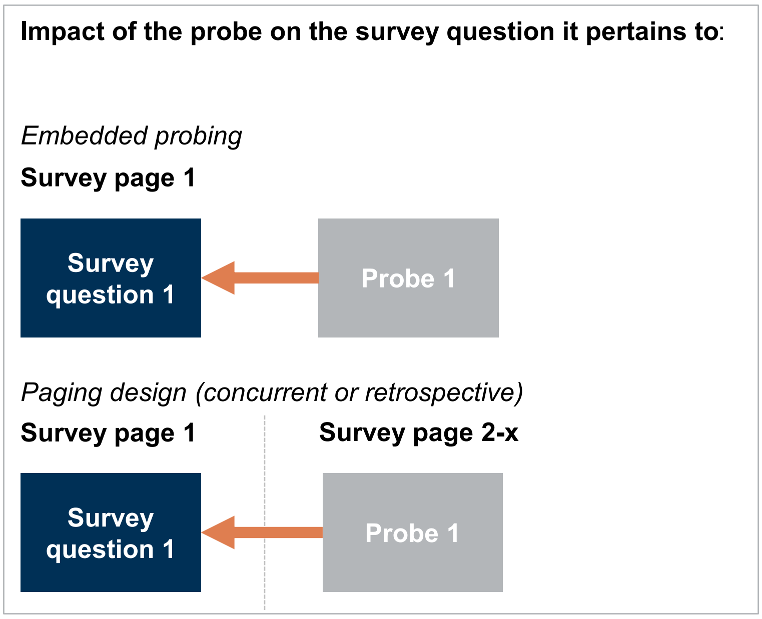


Figure 4.7. Impact of the probe on the survey question it pertains to

In the scenario of embedded probing, it is immediately apparent that adding a probe to a survey question constitutes a higher *response burden* (Yan & Williams, 2022) than only asking the survey question. Luebker (2021) demonstrated that respondents who received an embedded probe were more prone to item nonresponse than respondents who only received the survey question. In a study that examined a four-item grid, embedded probing increased item nonresponse for two items (Neuert & Lenzner, 2023). However, even if a probe is presented on a subsequent survey page, respondents will probably consider it an additional subtask of the survey question and perceive an increase in response burden. In Luebker's study, both embedded and concurrent probing resulted in increased survey break-offs. His findings correspond with research on open-ended questions in web surveys, which are associated with increased break-offs and decreased respondent motivation (Galesic, 2006; Peytchev, 2009). In studies including several open-ended probes, the increase in survey break-off may mainly be caused by the first probes, as a study containing either 13 or 21 probes found no difference in survey break-off between conditions (Neuert & Lenzner, 2021). A study examining the effect of closed probes on survey break-off revealed no significant differences between respondents who received probes and those who were only presented the survey questions; however, the study was not a randomized experiment but involved two rounds of a survey (Scanlon, 2019).

Besides increased response burden and consequences on nonresponse, implementing probes may impact the measurement of the survey question under examination through two mechanisms. First, when the survey question and probe are presented together on one page, respondents may choose to first read both the survey question and probe and then answer the survey question. In this scenario, respondents confront themselves with the probe before responding to the survey question. In consequence, respondents may process the survey question differently than they would if they were not simultaneously considering the probe. In this case, *reactivity* (i.e., Russo et al., 1989) potentially impacts survey response behaviour.

However, even when a probe is embedded on the same page as the survey question, respondents may process and respond to the survey question and probe consecutively. When a probe is presented on the survey page following the survey question (concurrent probing), the tasks of answering the survey question and probe must be carried out consecutively. In both cases, the cognitive task specific to probing is that of retrospection, meaning that respondents must recall the thoughts they had while answering the survey question. The previous section discussed that respondents might not remember the reasons for their survey response (*memory errors*) and might fill this memory gap by simply creating a probe response that is coherent with their chosen survey response. However, the reverse effect is also conceivable, with respondents choosing to make their survey response coherent with their probe response. Thus, the second mechanism that could impact survey response behaviour to the question being probed is based on the *maxim of relation*. Probe responses may be incoherent to survey responses because the differences in the response process between open-ended and closed questions (Schuman & Presser, 1979) lead respondents to consider additional or different aspects of a question during probing than they did while answering the survey question or respondents may come to a different evaluation as to which retrieved information should be relevant to their judgment. The maxim of relation requires that probe and survey responses do not contradict each other. If respondents reconsider their survey responses, this should be visible in an increase in answer changes to the survey question and in backtracking to the survey question when probes are asked concurrently. Eye-tracking could examine how respondents process survey questions and probes on the same page. Navigational paradata could capture backtracking and answer changes to the survey question in a concurrent design.

Two previous studies support the notion that implementing probes impacts survey response behaviour to the questions the probes pertain to. First evidence of such an effect was reported by Couper (2013, study 2), in an experiment in which a multi-item scale on attitudes towards immigrants was presented with one item per screen and an embedded probe accompanied each item. He reported a small but significant shift in means when embedding probes compared to a control group with no probes. In his study, respondents who received probes reported more positive attitudes towards immigrants.

A study by Fowler and Willis (2020) examined survey response behaviour to a multi-item battery on neighbourhood walkability, with probes being asked directly following the items or at the very end of the questionnaire. They found a small but significant change in the overall means of the item battery when probes were implemented concurrently. Unfortunately, this result remains highly inconclusive. For one, the authors did not report whether respondents had the opportunity to backtrack to previous survey pages to change their responses, which would be the only possible explanation for the change in means in a successfully randomized experiment. However, secondly, the reported study was not a randomized experiment; instead, the conditions with concurrent and retrospective probing were fielded several weeks apart.

In summary, previous research has demonstrated effects of (open-ended) probes on the survey questions they pertain to. An increase in *response burden* has been empirically verified, in that embedded probing increases item nonresponse to the survey question and both embedded and concurrent probing increase survey break-off (Luebker, 2021). However, the effects of probes on the response behaviour to the survey questions under examination merit more research. When respondents respond to the survey question and probe sequentially (for instance, in concurrent probing), thinking about the probe may cause them to re-evaluate their survey response and change the chosen response option to align with the probe response (*maxim of relation*). Such an effect could be demonstrated if respondents were more likely to backtrack to the survey question and change their survey response when probes are implemented than when no probes are implemented (backward context effect). However, if respondents come to expect probes, they may carry out introspection with subsequent survey questions, resulting in reactivity. The following section discusses such effects.

The impact of probes on subsequent survey questions

First evidence of an effect of open-ended probes on subsequent survey questions is reported by Couper (2013, study 1), who employed the same multi-item inventory on attitudes towards immigrants as before but placed probes concurrently on separate survey pages between the items. Couper found a small but significant effect on response distributions in this setting as of the second item. Thus, the study indicates that reactivity may extend to subsequent survey questions when respondents who have answered one probe come to expect probes to subsequent items (see Figure 4.8).

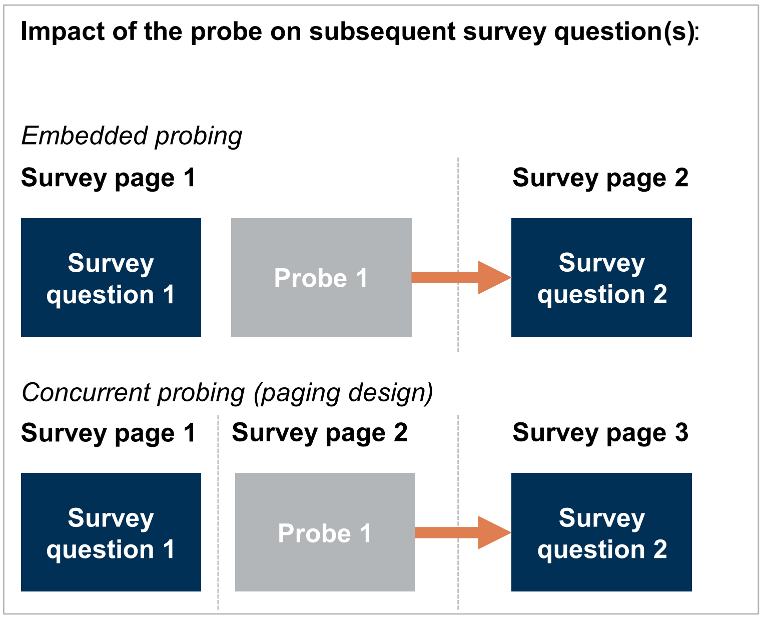


Figure 4.8. Impact of the probe on subsequent survey questions

Unfortunately, previous studies employing web probing give little indication of which types of effects on survey response behaviour to expect (both Couper, 2013, and Fowler & Willis, 2020, only report a shift in means). Knowles et al. (1992) argued that thinking about questions has consequences for question construal and that increased reflection on a topic makes a particular interpretation more salient, leading to a polarization of judgment. They postulated that later items within a measure (or items in a repeated measurement) show more extreme but more reliable and consistent responses. To examine this, the order of items within multi-item measures was randomized (Knowles, 1988), with later items showing higher reliability and more extreme answers. Notably, there was generally no visible effect on the mean values of these items. The studies demonstrated that increased reflection about survey questions influences cognitive processing and response to survey items and that these effects must not (necessarily) be visible by a simple comparison of means. Applied to the web probing context, the task of probing promotes a more systematic (Chen & Chaiken, 1999) and slower (Kahneman, 2012) processing of survey questions, which is likely to make more information accessible during the retrieval process (Tourangeau et al., 2000).

In summary, to extract the effects of probes on survey questions, experimental designs are required to compare survey questions accompanied by probes with identical survey questions without probes. Analyses should distinguish between effects on the survey questions the probes relate to and subsequent questions. Considering the ongoing discussion in cognitive pretesting that concurrent probing may interfere with the natural flow of the survey questions, understanding the mechanisms behind potential effects on survey response behaviour would be valuable. To this end, an examination of survey response behaviour that goes beyond an analysis of means to incorporate response styles is just as necessary as examining how respondents interact with survey questions in terms of survey navigation, such as backtracking, answer changes or response times.

4.2.3. The impact of probes on each other

Due to the endless possibilities of combining probes, examining how they impact each other poses the most comprehensive and complex range of scenarios. Common to all settings is that *response burden* is likely to increase with a rising number of (open-ended) probes, though even this effect is likely moderated by the specific context. In this section, *communicative maxims* and their effects on probe response content move to the centre of attention. The possible manifestations of these effects include the number and variety of themes, but also probe response styles such as the length of an answer and content density, and the prevalence of specific content. Two basic settings can be distinguished in which probes impact each other, those being (1) scenarios in which several probes pertain to the same survey question and (2) scenarios in which probes pertain to different survey questions.[8]

Several probes pertaining to the same survey question

The defining characteristic of settings where several probes follow one survey question is that each probe uses a different probing technique and focuses on a different aspect of the survey question and respondents' answering process (see Chapter 2.2). For instance, a category selection probe may ask respondents to explain their chosen survey response, followed by a comprehension probe on a particular term in the question text. There are three ways in which these probes can be presented (see Figure 4.9). In the embedded setting, the survey question and all related probes are presented together on the same survey page. Another possibility is to present the survey question on one screen, and all probes together on a separate screen, which we will refer to as a scrolling design. The third possibility is to employ a paging design, in which the survey question and each probe are presented on separate pages. The embedded and scrolling designs promote effects of probes on each other. In contrast, a paging design promotes that earlier-shown probes impact later probes (indicated by the arrows in Figure 4.9). As before, respondents may read and respond to probes on one survey page sequentially, and backward context effects are possible in a paging design when web survey programming permits backtracking.

Response burden should be highest in the embedded setting, as the perceived burden increases with the number of (survey and probing) questions on one survey page (Galesic, 2006; Peytchev, 2009). Regarding the perceived burden of answering the survey question, the embedded design should result in a higher level of item nonresponse to the survey question than the scrolling or paging design, in which the survey question is presented without any probes (see Luebker, 2021, for a study verifying this in a setting employing one probe only). Regarding the perceived response burden of answering the probes, the embedded and scrolling design should not differ greatly, as they both present several probes on one survey page. However, the perceived response burden of answering

[8] Of course, these settings can be combined. They will be discussed separately for the sake of simplicity.

the probes should be markedly lower in the paging design. Moreover, this effect should be stronger for the later-shown probe as response burden also increases with each preceding open-ended question (Galesic, 2006; Peytchev, 2009). Therefore, based on the mechanism of response burden, probe nonresponse should be lower, and the number of themes named should be higher in a paging design than in the embedded or scrolling designs, especially for the later probe.

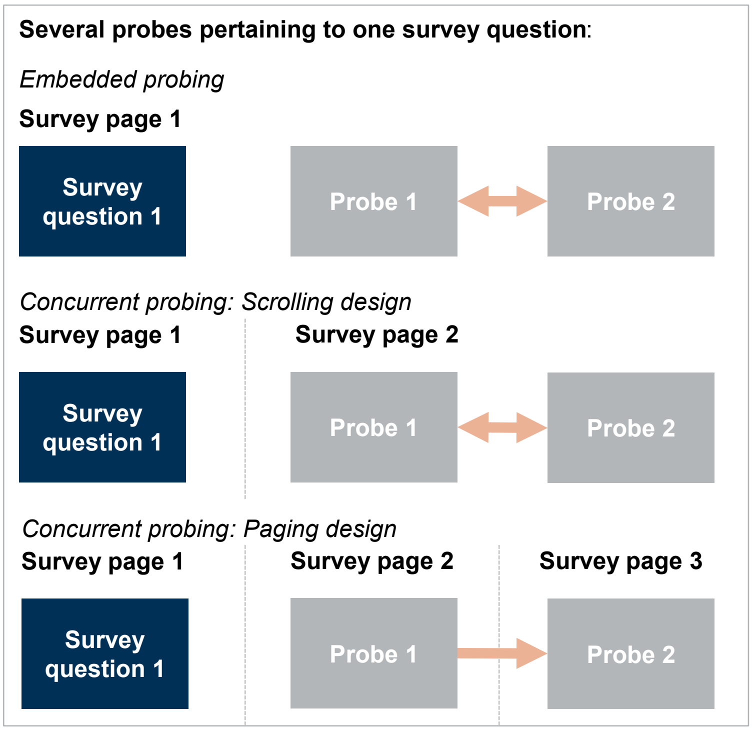


Figure 4.9. Impact of probes about one survey question on each other

On the one hand, the paging design decreases response burden, but on the other hand, it increases the distance between the survey question and later-shown probes. Therefore, the decrease in response burden in a paging design is accompanied by increased *memory errors*. With respondents less able to retrieve their response processes from short-term memory, they should become more likely to post-rationalize, drawing, for instance, on contextual information to generate their probe response. This is where *communicative maxims* become particularly relevant to research designs involving several probes. Respondents expect that each probe asks for new information from them. In return, they should provide new information in response to each probe (maxim of quantity), and the pieces of information should be coherent with each other and with the survey response (maxim of relation). When a paging design is employed, respondents who do not closely read the first-shown probe may believe that the second-shown probe breaches the *maxim of quantity* by asking the same information from them, resulting in a higher level of irritation in response to the second-shown probe (in contrast to an embedded or scrolling design, in which the respondents see both probes on one screen). Due to the increase in memory errors with each probe and the general need to give coherent probe responses to fulfil the *maxim of relation*, respondents should become more likely to generate their responses to later probes based on the responses they have given to previous probes. Thus, potentially a paging design results in more interpretable probe responses but with content of lower veracity.

Previous studies have lent support to these effects in web probing studies. In particular, the higher response burden of embedded and scrolling designs has been confirmed empirically (Meitinger et al., 2022; Neuert & Lenzner, 2023). Neuert and Lenzner (2023) compared the three settings for a single-item and a multi-item measure, each accompanied by two probes. They partially found a higher level of item nonresponse to the survey question in the embedded design compared to the paging design and a lower level of probe nonresponse in the paging design for the second-shown probe. Meitinger et al. (2022) found that the number of themes named was consistently lower for the second-shown probe in a scrolling compared to a paging design in a study that varied the sequence and the presentation of the probes.

In the first study that varied the sequence of probes about the same survey question, Meitinger et al. (2018) found that respondents were most likely to offer suitable probe response content when a category selection probe (rather than a comprehension or specific probe) was asked first. A follow-up study that varied both the order of two probes (category selection and specific) and the presentation of these probes (scrolling versus paging design) has since lent support to the application of both the maxim of quantity and relation regarding probe response content (Meitinger et al., 2022). For one, some respondents voiced irritation at the second-shown probe that they were being asked the same question twice, indicating that they felt the *maxim of quantity* was being breached. In all but one case, these respondents only provided content matching the category selection probe and no content that matched the specific probe. Thus, respondents generally avoided repeating information from one probe in the other.

The most striking finding in the study, however, is that it strongly indicates that the respondents partially suffer *memory errors* during probing and apply the *maxim of relation* to both probes, with marked effects on the response content to the second-shown probe. The survey question Meitinger et al. (2022) examined was, “The world would be a better place if people from other countries were more like the Germans”. The category selection probe asked them to explain their agreement or disagreement with the item. In contrast, the specific probe asked them to enumerate which countries they had been thinking about when answering the question. When the category selection probe was presented first, many respondents indicated ambiguity regarding the survey question, explaining that their response depends on which countries they consider. In contrast, when respondents first enumerated which countries they had in mind in answer to the specific probe, the share of these “it depends” answers to the category selection probe decreased significantly. The likely explanation for this effect is that some respondents did not consider specific countries while initially answering the survey question. When first presented with the specific probe, the probing task caused them to re-evaluate the survey question, giving a second thought to which countries would be logical to include. This led to a more concrete interpretation of the survey question, resulting in a lower share of “it depends” answers to the category selection probe when shown second. This finding aligns with the research by Knowles et al. (1992) that increased reflection on a topic cements a particular question interpretation, so respondents voice more marked opinions in later responses. Regarding context effects in web probing, it supports the idea that respondents’ interpretation of a survey question continues to evolve when they answer probes and that the answers to preceding probes may impact subsequent probes.

Probes pertaining to different survey questions

Compared to when several probes pertain to the same survey questions, probes about different survey questions offer an even wider variety of settings. Probing techniques may or may not vary across probes, and the survey questions may or may not be topically related. When probes are presented retrospectively, they are likely to directly follow each other, potentially increasing effects, whereas in embedded and concurrent settings, the survey questions they pertain to are interspersed between the probes (see Figure 4.10).

Again, one general effect to be expected is that *response burden* increases with the total number of open-ended probes (Galesic, 2006; Peytchev, 2009). In a study that employed concurrent probes and randomized the order of survey questions, Behr et al. (2012b) demonstrated that the share of substantive probe responses decreased between the first- and sixth-shown probe. Interestingly, the respondents who continued to provide substantive responses gave slightly longer responses to later probes. Supporting these findings, Neuert and Lenzner (2021) found that probe nonresponse increased in later-shown probes. In summary, perceived response burden increases with the number of preceding probes, resulting in a lower share of interpretable probe response content. However, respondents who continue to provide high-quality responses tend to give longer answers.

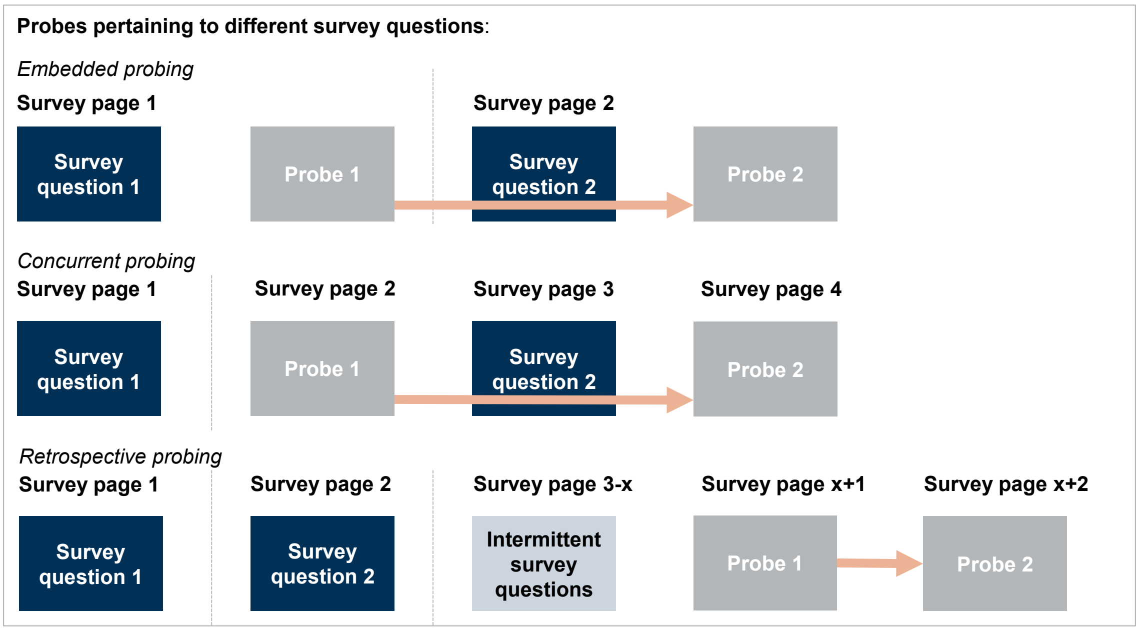


Figure 4.10. Impact of probes about different survey questions on each other

Another general effect established can best be described as an unwanted “learning” or *priming effect*. Behr, Bandilla, Kaczmirek, and Braun (2014) varied the number of category selection probes respondents received before a specific probe. Participants were most likely to provide mismatching probe response content when exposed to a high number of previous category selection probes and the visual design of a later-shown specific probe was kept consistent with that of the previous probes. Thus, while these respondents were motivated to provide substantive answers, they took shortcuts while reading the probing question, falsely assuming they were already familiar with the task.

Effects of response burden and priming due to visual cues are likely to apply regardless of whether survey questions are topically related. Potentially, probes relating to survey questions that share an overarching topic are subject to additional context effects. Settings that vary the order of the survey questions, and ultimately the order of related probes, are ideal for examining the potential effects of probes on each other. The challenge in examining these settings is that the order of related survey questions may impact the responses to the survey questions, as has been demonstrated for general and specific (i.e., Schwarz & Strack, 1991) and attitude and behaviour questions (i.e., Budd, 1987). The analysis must therefore distinguish between order effects on survey response, the effects of the survey questions (and survey response) on the probes, and the effects of the probes on each other. However, the same communicative maxims of relation and quantity apply.

Based on the *maxim of relation*, the content of probe responses should be coherent with the survey response. Consequently, shifts in response behaviour due to question order should be mirrored in shifts in probe response content (i.e., Bishop, 1992). For instance, in cases in which a change in the presentation of questions on life and relationship satisfaction leads to an increase in the correlation of the responses to the two survey questions, a probe asking about the aspects included in the judgment on life satisfaction should be more likely to include relationship satisfaction. Using the example of behaviour and attitude questions, attitude-behaviour consistency is higher when the behaviour question is asked first, especially when respondents establish a normative principle between the questions (Smith, 1992). Probe responses should mirror this increase in consistency. However, they may also be used to explain ‘inconsistent’ survey response behaviour, for instance when a respondent first expressed a strongly condemning attitude towards a topic but admits to engaging in this behaviour in a later survey question.

It quickly becomes apparent that, especially when two questions on the same topic are asked in a survey, respondents may include information related to the second survey question and probe in response to the first-shown probe already. For instance, in answer to a probe about a behaviour question, it seems logical that some respondents will include their attitude towards that behaviour in their probe response. In this case, the effects of the *maxim of quantity* must be discussed. If respondents are subsequently presented with an attitude question and probe on the same topic, they may perceive the second survey question and probe as repetitive. This may result in higher probe nonresponse, overt signs of irritation and fewer themes in later-shown probes.

In summary, previous research has demonstrated effects of preceding probes on later probes in terms of higher response burden, resulting in an increase in probe nonresponse and a decrease in the number of themes named (Behr et al., 2012b; Neuert & Lenzner, 2021), especially when probes are presented together on one survey page (Meitinger et al., 2022). Moreover, respondents highly exposed to the same probing technique are likely to assume that later probes require the same type of information from them (Behr, Bandilla et al., 2014). The effects of previous probe responses on later response content are complex and often incorporate the *maxim of relation*. An initial probe may cement a specific question interpretation, which will be the basis of the following probe response. First studies have found such effects for probes relating to the same survey question (Meitinger et al., 2022). No studies have examined whether similar effects exist when probes relate to different survey questions with an overarching topic. Moreover, the effects of the *maxim of quantity* remain under-researched, for instance, how respondents react when they have already provided content that is the topic of later-shown probes.

4.3. Research questions and objectives of the empirical studies

The preceding section proposed a psychological model of context effects in web probing organized around three main directions of effects: the effects of survey questions on probe responses, the effects of probes on survey responses, and the effects of probes on other probe responses. Previous research has established the existence of several effects in each of these areas. For example, that implementing open-ended probes increases *response burden* has been demonstrated in diverse settings (i.e., Behr et al., 2012b; Luebker, 2021; Neuert & Lenzner, 2023). However, effects associated with the probing-specific cognitive tasks of intro- and retrospection and communicative maxims remain under-researched. The following three research questions are organized around the main direction of context effects in web probing, each tackling the most urgent under-researched area:

Research Question 1: How do intermittent survey questions impact probe responses?

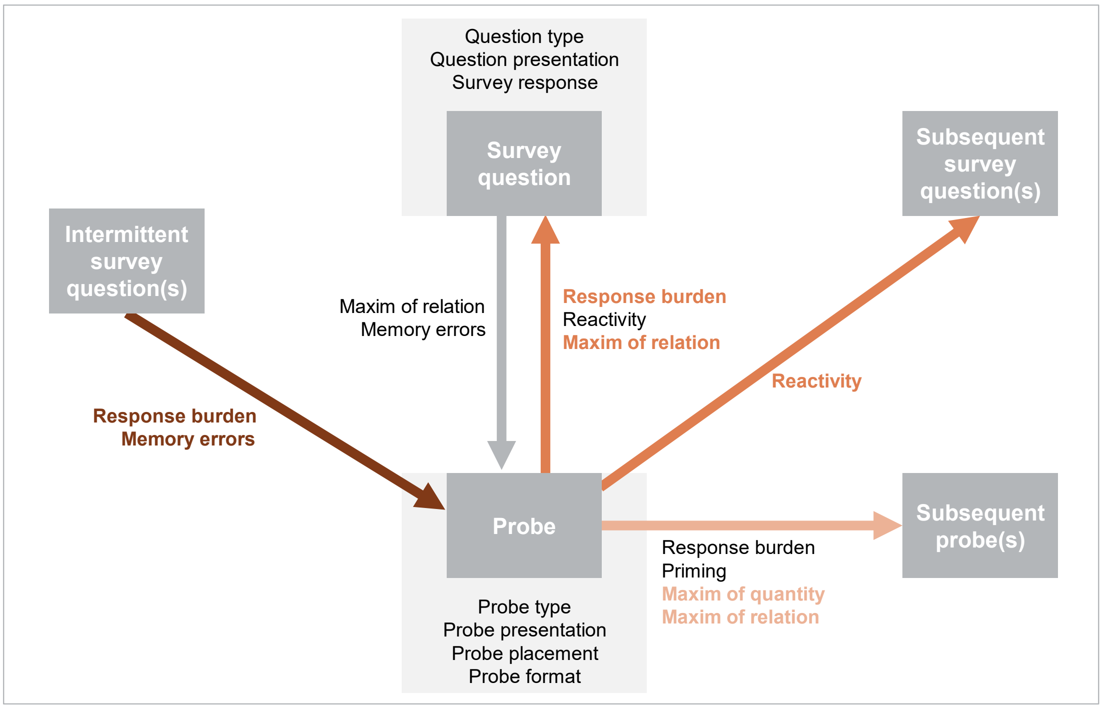
Does retrospective probing increase the perceived *response burden* of answering probes, and does it promote *memory errors* in probe response content?

Research Question 2: How does embedding probes concurrently impact surrounding survey questions?

First, does concurrent probing increase *response burden* and lead to re-evaluating the survey question the probe pertains to (*maxim of relation*)? Secondly, does it impact the processing of and response to subsequent survey questions (*reactivity*)?

Research Question 3: How do probes pertaining to different survey questions with an overarching topic impact each other?

Do the communicative *maxims of relation and quantity* impact how respondents answer probing questions when the order of the survey questions is varied?



Note. The dark orange arrow indicates the focus of study 1; the medium orange arrows are the focus of study 2; the light orange arrow is the focus of study 3.

Figure 4.11. Focus of the empirical studies

These research questions were examined in three empirical web probing studies. Figure 4.11 depicts the examined effects, with the shade of the orange arrows highlighting the direction(s) of effects and which underlying mechanisms will be examined in which study. The following sections summarize the studies' research questions and methods.

4.3.1. Study 1: The impact of intermittent survey questions on probes

The primary focus of the first study was to examine whether intermittent survey questions impact probe responses in retrospective probe placement. To this end, the study compared the perceived response burden and probe response quality of concurrent and retrospective probes pertaining to one general and two specific domains of quality of life. The study postulated retrospective probing would increase perceived response burden and memory errors. Perceived *response burden* was measured using the time spent reading and responding to a probe, and in attempts to leave a probe unanswered (activating a motivational prompt). The prevalence of *memory errors* was examined in terms of unconditional and conditional effects. The unconditional effect hypothesized an increase in non-substantive probe response content and a decrease in the number of themes in the retrospective condition. The conditional effect assumed that the topics of the specific domains would be named more often in response to the probe on the general domain in the retrospective condition.

Researchers have recently argued that closed or semi-open probe formats may decrease response burden and increase probe response quality (Scanlon, 2019, 2020). Therefore, the second focus of the study was to examine whether employing probes with predefined response options could decrease the assumed adverse effects of retrospective probe placement.

The study employed a 2x2 (probe placement x probe format) randomized web experiment with *N* = 2,184 respondents in Germany in November and December 2020. Response times and survey navigation data were collected using a client-side paradata script (UCSP; Kaczmirek & Neubarth, 2007). One general (life satisfaction) and two specific (relationship satisfaction and subjective health) domains of quality of life were accompanied by open-ended probes or probes with predefined response options placed concurrently or retrospectively. Probe responses were coded based on existing coding schemes and augmented using an inductive approach.

4.3.2. Study 2: The impact of probes on survey questions

The second study focused on the effect of implementing open-ended probes on the surrounding survey questions. The study distinguished between the effects of implementing probes on the overall response burden, the impact on the survey questions the probes pertain to, and the impact on subsequent survey questions. Overall *response burden* was gauged as the likelihood of survey break-off during the reported experiment. Based on the *maxim of relation*, the study hypothesized that respondents who were asked open-ended probes were more likely to backtrack to previous survey questions and change their survey responses than respondents who received no probes. The effects of implementing probes on subsequent survey questions through *reactivity* were examined using the measures of item nonresponse, response distributions, non-differentiation, extreme responding, and response times.

The study employed a randomized experiment comprising six single- and multi-item survey measures. Respondents were randomly assigned to a condition in which they received a concurrent open-ended probe after each of the six survey instruments or to a condition that included no probes. The web survey was fielded in November and December 2020 in Germany with a sample of *N* = 2,200 respondents from a non-probability online panel. Response times and survey navigation data were collected using a client-side paradata script (UCSP; Kaczmirek & Neubarth, 2007).

4.3.3. Study 3: The impact of question order on probes

The third study examined the effects of the order of two survey questions and concurrent probes on probe response content. More precisely, it examined the application of the *maxim of relation* in terms of the consistency of probe responses to one another and the *maxim of quantity* in the prevalence of probe response content. To this end, an attitude and behaviour question, each followed by a concurrent probe, were implemented in a web survey, and the order of the survey questions and, ultimately, probes was randomized. Attitude-behaviour consistency is generally higher when the behaviour question is asked first (Budd, 1987) as respondents are likely to align their attitudinal responses with self-reports on behaviour, especially when they establish a normative principle between two questions (Smith, 1992), such as for delinquent behaviour. Based on the *maxim of quantity*, the likelihood of mentioning behaviour-, attitude-, and problem-related content was examined as a function of question order. This effect was examined with two competing hypotheses, namely that response content was generally less likely to be named in answer to the second-shown probe (unconditional effect) or that it was only less likely to be named if respondents had already referred to their behaviour or attitude in response to the first-shown probe (conditional effect). Based on the *maxim of relation*, the consistency of attitude- and behaviour-related probe response content was postulated to increase when the behaviour question was asked first. Finally, this study took a cross-cultural perspective and examined the order effects in a U.S. and German sample.

The web experiment was conducted in July and August 2018 with *N* = 333 respondents from non-probability samples in Germany and the United States. Respondents were randomly assigned to a condition where they first received the behaviour survey question and probe, followed by the attitude survey question and probe, or vice versa. Responses to both probes were coded as to whether they contained behaviour-related or attitude-related content and whether respondents expressed difficulties with the respective survey question. Probe responses were coded as consistent when respondents either reported that they had committed the delinquency and voiced a more lenient attitude towards it or reported that they had never committed it and expressed absolute condemnation for the behaviour.

4.3.4. Publication status of the studies

The three studies have been published or submitted for publication as:

1. Hadler, Patricia (submitted). Response burden and response quality in web probing: An experiment on the effects of probe placement and format. *Survey Research Methods*.

2. Hadler, Patricia (2023). The effects of open-ended probes on closed survey questions in web surveys. *Sociological Methods & Research*, Online First, 1-34. DOI: 10.1177/00491241231176846.

3. Hadler, Patricia (2021). Question order effects in cross-cultural web probing: Pretesting behavior and attitude questions. *Social Science Computer Review, 39*(6), 1292-1312. DOI: 10.1177/0894439321992779.

5. STUDY 1: THE IMPACT OF INTERMITTENT SURVEY QUESTIONS

A version of this chapter has been submitted to *Survey Research Methods* as

Hadler, Patricia. Response burden and response quality in web probing: An experiment on the effects of probe placement and format.

5.1. 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). 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).

A fundamental research design decision when implementing probes is probe ***placement***, or when to ask the probing question (Willis, 2005, p. 51f.). 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.

5.2. 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).

5.2.1. 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 recap 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. 51f.), 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” approaches (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.

5.2.2. 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 and be 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.

5.3. 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.

Table 5.1. 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

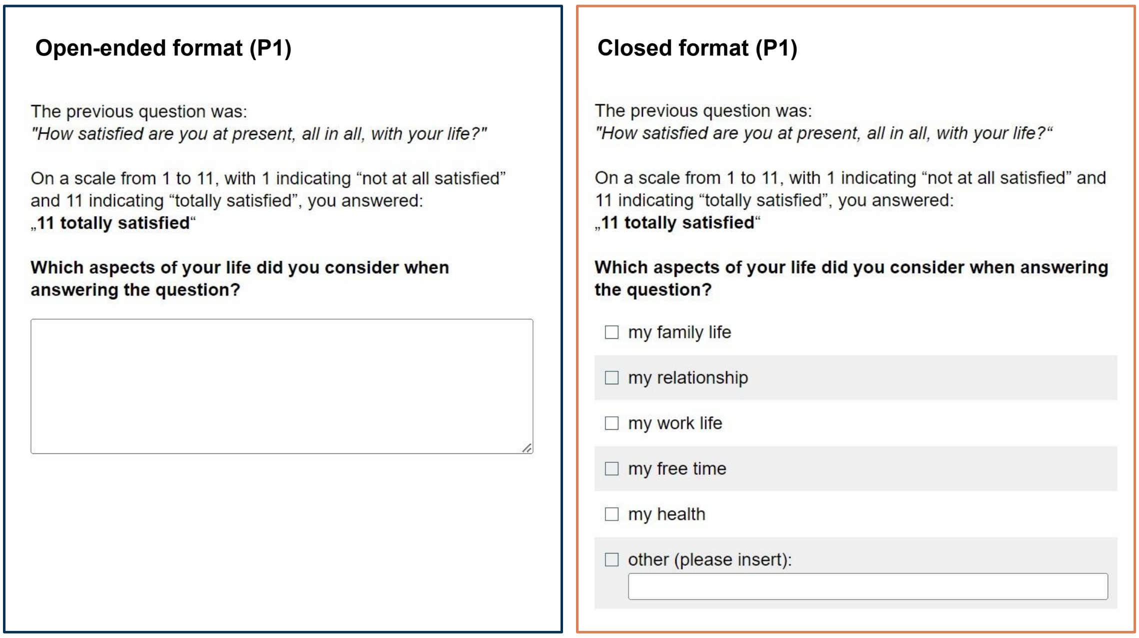
H3f: … the use of memory cues from intermittent survey questions

Regarding the third hypothesis on the ***moderating effects of probe format***, I hypothesize that the adverse 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 regarding the parameters mentioned above (H3a to H3f). Thus, an interaction effect of probe placement and format is assumed. Table 5.1 summarizes the hypotheses.

5.4. Method

5.4.1. 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 Figure 5.1), 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 5.2).



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):

Figure 5.1. Probe format

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 4,994 respondents (36.2%) 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 4,386 respondents who completed the questionnaire, 2,184 were part of the current experiment. The sample included quotas to depict the German online population in terms of gender (male, female)[9] 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.

Table 5.2. Experimental conditions

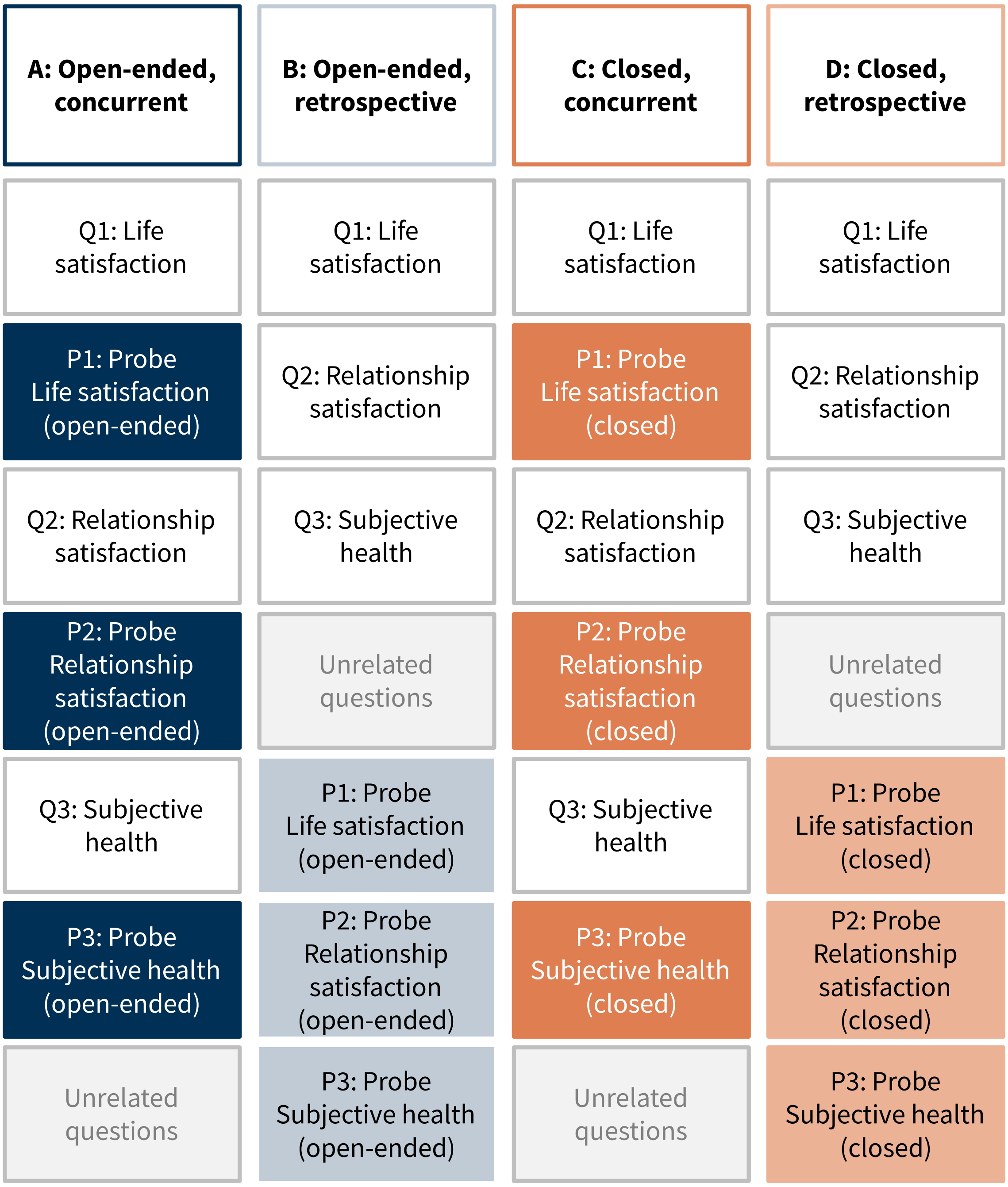
	Probe format	
Probe placement	Open (Open-ended text field)	Closed (Predefined response options)
Concurrent	A: Open-ended, concurrent	C: Closed, concurrent
Retrospective	B: Open-ended, retrospective	D: Closed, retrospective

The reported study was placed towards the beginning of the survey, after the quota-relevant questions and one other experiment. 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. Then the three probes immediately followed each other (see Figure 5.2).

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, &

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

Schaurer, 2020), respondents were informed about the collection and use of client-side paradata on the welcome page of the survey.



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)

Figure 5.2. Study design

5.4.2. 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” (Bruin et al., 1996). There were no significant differences in response 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.”).

5.4.3. 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 relationship 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 .948 (P1), .856 (P2), and .921 (P3). The author and a student assistant independently coded the substantive responses. For the predefined categories, an intercoder reliability of .980 to 1.000 (P1), .832 to .987 (P2) and .867 to 1.000 (P3) was reached. For the additional themes that emerged, Cohen's Kappa ranged from .896 to .992 (P1), .841 to .930 (P2) and .778 to .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.

5.4.4. 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 5.6% and 9.3% 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). 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.

5.5. Results

5.5.1. 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, 1,308 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 5.3). Gender, age, and device were significant covariates, whereas education was not.[10] Figure 5.3 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.

[10] MANCOVA results with respect to the covariates (1) gender: *Wilks-Lambda* = 0.99; $F_{(3,1298)}$ = 5.40; p = .001; η^2 = .01; (2) age: *Wilks-Lambda* = 0.87; $F_{(3,1289)}$ = 67.29; p < .001; η^2 = .13; (3) education: *Wilks-Lambda* = 1.00; $F_{(3,1298)}$ = .88; p = .453 n.s.; (4) device used: *Wilks-Lambda* = 0.96; $F_{(3,1298)}$ = 18.77; p < .001; η^2 = .04

Table 5.3. Probe response times, MANCOVAs

	Response latency				Time spent answering			
N	1,308				1,332			
Main predictors	Wilk's λ	$F_{(3,1298)}$	*p*	η^2	Wilk's λ	$F_{(3,1222)}$	*p*	η^2
Placement*format	0.98	9.62	< .001	.02	0.99	2.38	.068	-
Probe placement	0.94	27.91	< .001	.06	1.00	1.08	.356	-
Probe format	0.67	215.66	< .001	.33	0.59	306.77	< .001	.41
Between-subjects effects		$F_{(1,1300)}$	*p*	η^2		$F_{(1,1324)}$	*p*	η^2
Placement*format								
P1		1.29	.255	-		-	-	-
P2		11.18	.001	.01		-	-	-
P3		10.65	.001	.01		-	-	-
Probe placement								
P1		52.48	< .001	.04		-	-	-
P2		50.17	< .001	.04		-	-	-
P3		62.25	< .001	.05		-	-	-
Probe format								
P1		33.23	< .001	.02		760.39	< .001	.37
P2		294.01	< .001	.18		394.00	< .001	.23
P3		540.55	< .001	.29		488.88	< .001	.27

Regarding the *time spent answering*, 1,332 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 5.3), lending no support to H1b or H3b. Again, the probe format exerted a strong and significant influence. Age was the only significant covariate.[11] A MANCOVA based on the alternative response time outlier exclusion confirmed the overall and between-subjects effects. The lower row of Figure 5.3 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

[11] MANCOVA results with respect to the covariates (1) gender: *Wilks-Lambda* = 1.00; $F_{(3,1322)} = 2.09$; *p* = .010; n.s.; (2) age: *Wilks-Lambda* = 0.96; $F_{(3,1322)} = 16.64$; *p* < .001; $\eta^2 = .04$; (3) education: *Wilks-Lambda* = 1.00; $F_{(3,1322)} = 1.84$; *p* = .138 n.s.; (4) device used: *Wilks-Lambda* = 1.00; $F_{(3,1322)} = 1.56$; *p* = .197; n.s

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.

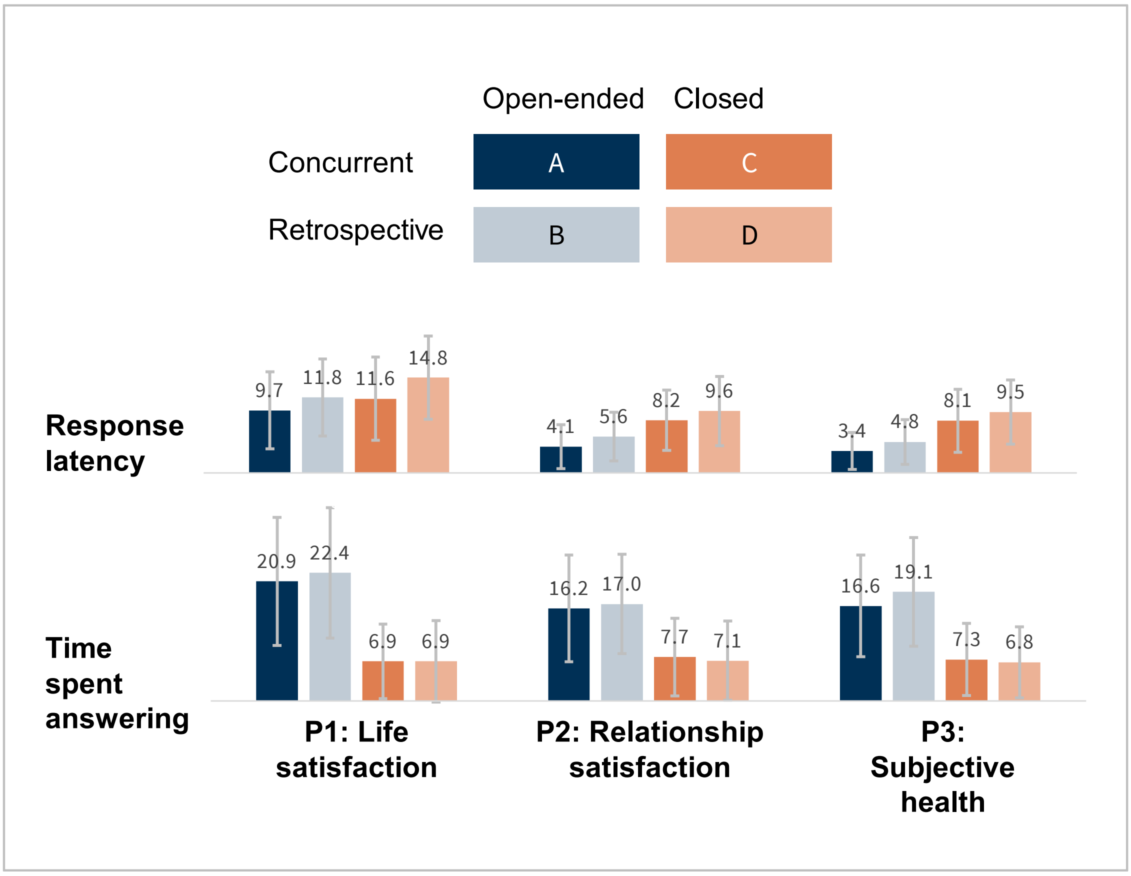


Figure 5.3. Probe response times (mean and standard deviation)

Motivational prompts. In total, only 69 (3.2%) 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: 3.1%; retrospective: 3.2%; $\chi^2(1) = .010$; $p = .921$), whereas open-ended probes were significantly more likely to be associated with activating prompts than probes with predefined response options (open-ended: 5.2%; closed: 1.1%; $\chi^2(1) = 30.733$; $p < .001$).

5.5.2. 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).

Non-substantive probe responses. Table 5.4 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 19.6% and 34.5%) than for probes in the check-all-that-apply format (between 1.5% and 3.5%). 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.

Table 5.4. Non-substantive probe responses, binary logistic regressions

	P1: Life satisfaction	**P2: Relationship satisfaction**	**P3: Subjective health**
N (Basis: all probe responses)	2,181	2,181	2,181
	% (n)	% (n)	% (n)
A: Open-ended, concurrent	19.6% (106)	31.0% (168)	28.2% (153)
B: Open-ended, retrospective	24.8% (135)	34.5% (188)	30.8% (168)
C: Closed, concurrent	1.6% (9)	1.8% (10)	1.5% (8)
D: Closed, retrospective	3.5% (19)	2.9% (16)	2.4% (13)
Binary logistic regression	OR	OR	OR
Placement*format	.61	.70	.69
Probe placement	1.69*	1.36	1.33
Probe format	.08***	.05***	.04***
Model χ^2 (7)	258.35***	442.83***	428.35***
Correct classification (%)	87.7	82.5	84.7
Nagelkerke R^2	.213	.304	.308

**p* < .05; ***p* < .01; ****p* < .001

Mean number of themes. Table 5.5 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 5.5. Mean number of themes, descriptive results

	P1: Life satisfaction	**P2: Relationship satisfaction**	**P3: Subjective health**
N (Basis: substantive probe responses)	1,915	1,802	1,842
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)

After excluding non-substantive responses, 1,610 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 5.6), 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.[12] 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).

Table 5.6. Mean number of themes, MANCOVA

		Mean number of themes		
N		1,610		
Main predictors	Wilk's λ	$F_{(3,1600)}$	*p*	η^2
Placement*format	1.00	0.89	.443	-
Probe placement	1.00	1.00	.391	-
Probe format	0.87	79.46	< .001	.13
Between-subjects effects for significant predictors				
Probe format		$F_{(1,1602)}$	*p*	η^2
P1		0.40	.529	-
P2		90.24	< .001	.05
P3		160.97	< .001	.09

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

[12] MANCOVA results with respect to the covariates (1) gender: *Wilks-Lambda* = 0.99; $F_{(3,1600)} = 3.10$; $p = .027$; $\eta^2 = .01$; (2) age: *Wilks-Lambda* = 0.97; $F_{(3,1600)} = 15.82$; $p < .001$; $\eta^2 = .03$; (3) education: *Wilks-Lambda* = 1.00; $F_{(3,1298)} = 9.22$; $p < .001$; $\eta^2 = .02$; (4) device used: *Wilks-Lambda* = 1.00; $F_{(3,1600)} = .40$; $p = .750$; n.s.

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 5.7 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.

Table 5.7. Themes "relationship" and "health", binary logistic regressions

	Relationship	**Health**
N (Basis: substantive probe responses)		
	% (n)	% (n)
A: Open-ended, concurrent	16.1% (70)	40.6% (177)
B: Open-ended, retrospective	25.9% (106)	38.0% (156)
C: Closed, concurrent	48.9% (263)	63.6% (342)
D: Closed, retrospective	47.6% (253)	58.4% (310)
Binary logistic regression	OR	OR
Placement*format	.474***	.923
Probe placement	.732**	1.189
Probe format	.264***	.419***
Model χ^2 (7)	232.63***	115.23***
Correct classification (%)	67.6	61.2
Nagelkerke R^2	.157	.078

**p* < .05; ***p* < .01; ****p* < .001

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 [.31, .72]), as well as significant main effects of predictors. In the open-ended condition, only 16.1% of respondents named the theme "relationship" when the probe was asked concurrently, whereas 25.9% 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: 48.9%; retrospective: 47.6%).

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* = .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”.

5.6. 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 confirmed. 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.

Table 5.8. Study 1: 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

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 5.8.

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 lower 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.

5.7. Appendix Study 1

A.1. Sample composition

	A: Open-ended, concurrent	**B: Open-ended, retrospective**	**C: Closed, concurrent**	**D: Closed, retrospective**	**Significance level**
Gender	% (*n*)	% (*n*)	% (*n*)	% (*n*)	
Male	47.0% (255)	51.6% (281)	51.2% (280)	51.5% (283)	$\chi^2_{(3)} = 3.10$; $p = .376$
Female/ non-binary	53.0% (287)	48.4% (264)	48.8% (267)	48.5% (267)	
Education					
Low	41.1% (223)	50.5% (275)	47.7% (261)	46.0% (253)	$\chi^2_{(6)} = 11.08$; $p = .086$
High	58.7% (318)	49.4% (269)	52.1% (285)	54.0% (297)	
Unknown	0.2% (1)	0.2% (1)	0.2% (1)	0.0% (0)	
Device used					
PC/Laptop	61.1% (331)	63.9% (348)	65.1% (356)	62.4% (343)	$\chi^2_{(9)} = 5.03$; $p = .832$
Tablet	8.3% (45)	8.8% (48)	9.1% (50)	8.0% (44)	
Smartphone	30.4% (165)	27.0% (147)	25.4% (139)	29.5% (162)	
Other/ unknown	0.2% (1)	0.4% (2)	0.4% (2)	0.2% (1)	
Age	mean (std)	mean (std)	mean (std)	mean (std)	
Mean age (in years)	43.6 (14.76)	45.4 (15.00)	45.6 (14.90)	45.1 (15.28)	$F_{(3,2180)} = 1.97$; $p = .117$

A.2. Questionnaire

Q1. The following question is about your general life satisfaction. How satisfied are you at present, all in all, with your life? *[Nun geht es um Ihre allgemeine Lebenszufriedenheit. Wie zufrieden sind Sie gegenwärtig, alles in allem, mit Ihrem Leben?]*

Response scale: 1 not at all satisfied, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11 totally satisfied, I do not want to answer *[1 überhaupt nicht zufrieden, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11 völlig zufrieden, das möchte ich nicht beantworten]*

P1a [open]. Specific probe

The previous question / One of the previous questions was: [Q1]. On a scale from 1 to 11, with 1 indicating “not at all satisfied” and 11 indicating “totally satisfied”, you answered “[response to Q1]”. Which aspects of your life did you consider when answering the question? *[Die Frage soeben lautete: Q1. Auf einer Skala von 1 bis 11, wobei 1 "überhaupt nicht zufrieden" und 11 "völlig zufrieden" bedeutet, haben Sie den Wert „[response to Q1]“ angekreuzt. An welche Aspekte Ihres Lebens haben Sie beim Beantworten der Frage gedacht?]*

P1b [closed]. Specific probe

Probe question identical to P1a.

Response options [multiple choice]: my health, my work life, my family life, my free time, my relationship, other (please insert): [open-ended text field] *[meine Gesundheit, mein Arbeitsleben, mein Familienleben, meine Freizeit, mein Beziehungsleben, anderes (bitte angeben): [open-ended text field]]*

Q2. Please think about your current relationship (marriage or partnership). How satisfied are you at present with your relationship? *[Denken Sie bitte einmal an Ihre partnerschaftliche Beziehung (Ehe oder Freund/in). Wie zufrieden sind Sie zurzeit mit Ihrer Partnerschaft?]*

Response scale: 1 not at all satisfied, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11 totally satisfied, I do not want to answer *[1 überhaupt nicht zufrieden, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11 völlig zufrieden, das möchte ich nicht beantworten]*

P2a [open]. Specific probe

The previous question / One of the previous questions was: [Q2]. On a scale from 1 to 11, with 1 indicating “not at all satisfied” and 11 indicating “totally satisfied”, you answered “[response to Q2]”. Which aspects of your relationship did you consider when answering the question? *[Die Frage soeben lautete: Q2. Auf einer Skala von 1 bis 11, wobei 1 "überhaupt nicht zufrieden" und 11 "völlig zufrieden" bedeutet, haben Sie den Wert „[response to Q2]“ angekreuzt. An welche Aspekte Ihrer Partnerschaft haben Sie beim Beantworten der Frage gedacht?]*

P2 [closed]. Specific probe

Probe question identical to P2 [open-ended].

Response options [multiple choice]: the commitment in my relationship, the freedom within the relationship, my relationship status, faithfulness, mutual sexuality, mutual free-time activities, other (please insert): [open-ended text field] *[An die Verbindlichkeit meiner Beziehung, das Commitment; An meine Freiräume in der Beziehung; An meinen Beziehungsstatus; An Treue; An die gemeinsame Sexualität; An gemeinsame Freizeitgestaltung und Aktivitäten; An anderes (bitte angeben): (open-ended text field)]*

Q3. How would you rate your overall health? *[Wie ist Ihr Gesundheitszustand im Allgemeinen?]*

Response scale: Very good, Good, Medium, Poor, Very poor *[Sehr gut, Gut, Mittelmäßig, Schlecht, Sehr schlecht]*

P3a [open]. Specific probe.

The previous question / One of the previous questions was: [Q3]. You answered "[response to Q3]". Which aspects of your health did you consider when answering the question? *[Die Frage soeben lautete: „Q3" Ihre Antwort lautete: „[Response Q3]". An welche Aspekte Ihrer Gesundheit haben Sie beim Beantworten der Frage gedacht?]*

P3b [closed]. Specific probe.

Probe question identical to P3a.

Response options [multiple choice]: Whether I frequently experienced pain (i.e., headaches, pain in the back or limbs); Whether or how often I visited a doctor; My emotional state; Whether or how much medication I took; Whether and which illnesses I had lately; My mental capacity; Other (please insert): [open-ended text field] *[Ob ich häufig unter Schmerzen litt (bspw. Kopf-, Rücken- oder Gliederschmerzen); Ob bzw. wie oft ich einen Arzt/eine Ärztin aufgesucht habe; An mein emotionales Gleichgewicht; Ob bzw. wie viele Medikamente ich eingenommen habe; ob bzw. welche Krankheiten ich in letzter Zeit hatte; An meine geistigen Fähigkeiten; An anderes (bitte angeben): (open-ended text field)]*

A.3. Coding schemes (substantive codes)

Code name	Definition	Examples
P1. Life satisfaction		
Predefined themes		
Health	Reference to the physical or mental health of the respondent (NOT: health of family members or friends)	Health; health issues; mental health; feelings; depression; illness; chronic disease; wheelchair; addiction; weight
Work	Reference to respondent's work life, unemployment, retirement, or education (NOT: income or financial situation)	Work; job; career; new job; office; job is safe; unemployed; out of work; looking for job; home office; reduced working hours; school; apprenticeship; internship; education; retired
Family	Reference to the family, including health or well-being of family members, death or tragedy in the family, relatives or pets, and desire to have children	Family; how my family is doing; my family is healthy; my uncle died; parents; children; our dog; we want to have a baby
Leisure time	Reference to leisure time and activities, hobbies, friendships, travel, or sport	Free time; friends; social environment; social life; travel; holidays; sport; fitness; training; sport club; hobbies; volunteering; culture; the movies; going out
Relationship	Reference to partner or love life, including lack thereof	Boyfriend; girlfriend; husband; wife; spouse; relationship; sex; love life; love; single
Additional themes		
Financial situation	Reference to the financial situation, including salary, wealth, retirement, or pension	Finances; retirement pension; rent; income; wealth; assets; how much I make
Corona	Direct mention of Covid pandemic and/or its effect on everyday life	Corona; Covid; pandemic; restrictions due to Covid
Living situation	Reference to the living and housing situation	Living situation; neighbourhood; house; apartment; garden; the town; my home
Well-being	Reference to the respondent’s subjective well-being	Well-being; happiness; happy with my life; satisfied with life; sense of purpose

(Table A.3 continued)

Code name	**Definition**	**Examples**
Future and goals	Reference to the respondent's future plans or goals, worries about the future, but also success, independence and freedom	Future; prospects; goals; plan for the future; success; independence; freedom
Politics, society, environment	Reference to the political or societal environment, including global issues	Politics; society; environment; climate change; state of the world; economic situation; national security; gender debate
Private life	Respondent names "private life", which may encompass family, relationship, friendship and/or leisure time	Private life
Lifestyle, standard of living	Reference to the respondent's lifestyle and standard of living	Standard of living; lifestyle; living circumstances; quality of life; living in Germany
Loneliness and stress	Reference to loneliness or stress, worries or feelings of overburdened	Lonely; along; stress; worries; overstrained; bored; keeping me up at night
Other	Mention of further categories, including everyday routine, household, faith	Faith; everyday life; household duties; taking care of children
P2. Relationship satisfaction		
Predefined themes		
Relationship status	Reference to whether the respondent has a relationship	I don't have a partner; divorced; separated; widow; in a new relationship
Activities	Reference to mutual activities and how much time is spent together	Things we do together; holiday; time spent together; travel; going out; long-distance relationship; quality time; how much contact we have
Sexuality	Reference to sexuality	Sex; sex life; how often sex; intimacy; kissing; physical closeness
Faithfulness	Reference to faithfulness and monogamy, or lack thereof	I am seeing someone on the side; he cheated; faithful
Commitment	Reference to the commitment to the relationship, including emotional support, future planning, and standing side by side	Future plans; we have the same goals; moving in together; engagement; we are a team; on the same side; being there for each other; belong together; we can depend on each other; security

(Table A.3 continued)

Code name	**Definition**	**Examples**
Personal space	Reference to personal freedom within the relationship	Individual freedom; my partner lets me be free; I have my personal space; opportunity to grow
Additional themes		
Feeling of trust and happiness	Reference to trust or emotional closeness and feeling of happiness	Trust; familiarity; connectedness; warmth; feeling of happiness; he/she makes me happy; feeling of well-being when together
Harmony and arguments	Reference to how well the couple gets along, both positively and negatively	Get along perfectly; arguments; difference of opinion; criticism; relationship crisis; problems; good / bad atmosphere; harmony; stress due to outer circumstance
Feeling of love and affection	Reference to emotional attachment, love, being in love	Love; true love; in love; I feel loved; affection; emotional state; romance
Daily life	Reference to daily life as a couple and living together	Life together; living together; household; responsibility; distribution of chores; equality; daily life; daily routing; boredom; variety; spontaneity
Treatment and interaction	Reference to how the couple treats each other and interacts with each other	Considerateness; understanding; appreciation; how he treats me; respect; accepts me how I am; attention; empathetic; feel taken seriously
Communication and humour	Reference to communication within the relationship, including humor	Communication; discussions; honesty; listening to the other person; talking a lot; open and honest; exchange; problem solving; humor; fun; laugh together
Character and attitudes	Reference to the partner’s character and attitudes	Values; opinions; attitudes; sharing the same beliefs; things we have in common; mutual interests; character; change in character; our strengths complement each other; compatibility
Being family and friends	Reference to partner as part of one's family or the partner as a best friend	My wife is my family; friendship is the basis of a relationship; our marriage and children
Other	Reference to further distinct themes, including housing situation, working together, partner’s health, or Covid restrictions	Living space; living situation; finances; we work together; my husband is ill; I take care of my wife; Covid; age difference; what he looks like

(Table A.3 continued)

Code name	Definition	Examples
P3. Subjective health		
Predefined themes		
Illness	Reference to acute or chronic illnesses or operations	Chronic disease; Covid; the flu; cancer; I am sick; high blood pressure
Pain	Reference to acute or chronic pain, including pain location	In pain; back hurts; chronic headaches; migraine; muscle pain
Emotional Health	Reference to emotional health including mental health	Psychological health; mourning; depression; burn out; bipolar; mobbing; panic attacks; anxiety
Mental Fitness	Reference to mental capacity or fitness	Mental fitness; memory; mental state; mental training
Doctor visits	Reference to doctor visits or other medical specialists	Doctor appointment; check ups; sick days; waiting time
Medication	Reference to medication	Medication; pills; I don’t take meds; opioids; eight different pills
Additional themes		
Afflictions	Complaints or absence of complaints	Impairment; affliction; discomfort; health problems; minor ailments
Fitness and sports	Reference to physical fitness, activity or sport	Fitness; fit; physical fitness; sports; physical activity; training; yoga; tennis; moving a lot; I can’t walk; condition; energy
Age and independence	Reference to respondent's age, age-dependent health, mobility or independence	Age; symptoms of old age; good for my age; I am getting old; age-typical symptoms; compared to others my age; I can live independently; mobility; I can’t stand for long; rollator; menopause; losing hair
Weight, diet, alcohol, smoking	Reference to weight, diet, and health-related behaviours	Weight; I weigh too much; BMI; I want to loose weight; diet; nutrition; I don’t eat healthily; smoking; quit smoking; I drink too much; I could live more healthily
Body & psyche	Respondent refers to the unity of physical and psychological health	Mental and physical health; physical and psychological; emotional and bodily state or well-being
Well-being	Reference to general well-being, that is reference to how well one feels; vague whether it refers to physical or psychological factors	Well-being; how I feel; general feeling

(Table A.3 continued)

Code name	**Definition**	**Examples**
Stress	Reference to stress, lack of sleep, negative environmental influences such as work-related stress or risks	Too much stress; tired; exhaustion; fatigue; weakness; sleep; sleep problems; no time for myself; my environment is toxic; too much work; pain due to working from home
Physical health	Explicit reference to physical health (exclusion of or no mention of psychological health)	My body; physical health; how I am built; from head to toe
Handicap	Reference to a handicap or a handicap status	Handicap; physical handicap; multiple disabilities; degree of disability [with %]; deaf; blind
Other	Reference to other health-related themes, including immune system or injury	immune system; tendency to get ill; pregnancy; injured; broken bone; artificial joint

A.4. Distribution of predefined themes by experimental condition

(Basis: substantive responses)

	A: Open-ended, concurrent	**B: Open-ended, retrospective**	**C: Closed, concurrent**	**D: Closed, retrospective**
	% (n)	% (n)	% (n)	% (n)
P1: Life satisfaction				
Health	40.6% (177)^{c,d}	38.0% (156)^{c,d}	63.6% (342)^{a,b}	58.4% (310)^{a,b}
Work	39.2% (171)^{c,d}	43.2% (177)$^{c,(d)}$	53.3% (287)^{a,b}	51.4% (273)$^{a,(b)}$
Family	34.4% (150)^{c,d}	28.5% (117)^{c,d}	50.6% (272)^{a,b}	50.8% (270)^{a,b}
Leisure time	24.5% (107)^{c,d}	21.0% (86)^{c,d}	50.7% (273)^{a,b}	48.6% (258)^{a,b}
Relationship	16.1% (70)^{b,c,d}	25.9% (106)^{a,c,d}	48.9% (263)^{a,b}	47.6% (253)^{a,b}
P2: Relationship satisfaction				
Relationship status	31.0% (116)^{c,d}	32.2% (115)^{c,d}	62.0% (333)^{a,b}	56.4% (301)^{a,b}
Activities	10.2% (38)^{c,d}	9.5% (34)^{c,d}	44.7% (240)^{a,b}	44.2% (236)^{a,b}
Sexuality	9.4% (35)^{c,d}	10.4% (37)^{c,d}	34.1% (183)^{a,b}	34.5% (184)^{a,b}
Faithfulness	7.0% (26)^{c,d}	7.0% (25)^{c,d}	38.5% (207)$^{a,b,(d)}$	30.9% (165)$^{a,b,(c)}$
Commitment	15.5% (58)$^{(b),c,d}$	9.0% (32)$^{(a),c,d}$	31.3% (168)^{a,b}	30.1% (161)^{a,b}
Personal space	0.5% (2)^{c,d}	0.8% (3)^{c,d}	34.3% (184)^{a,b}	29.8% (159)^{a,b}
P3: Subjective health				
Illness	41.4% (161)^{c,d}	37.9% (143)^{c,d}	54.4% (293)^{a,b}	51.8% (278)^{a,b}
Pain	15.9% (62)^{c,d}	18.0% (68)^{c,d}	59.6% (321)^{a,b}	55.3% (297)^{a,b}
Emotional Health	6.9% (27)^{c,d}	6.4% (24)^{c,d}	43.6% (235)^{a,b}	42.6% (229)^{a,b}
Mental Fitness	0.8% (3)^{c,d}	0.5% (2)^{c,d}	30.8% (166)^{a,b}	27.6% (148)^{a,b}
Doctor visits	1.8% (7)^{c,d}	2.4% (9)^{c,d}	28.0% (151)^{a,b}	25.3% (136)^{a,b}
Medication	0.8% (3)^{c,d}	1.3% (5)^{c,d}	25.6% (138)^{a,b}	24.8% (133)^{a,b}

Note. Letters in superscript indicate significant differences as per chi-square test of independence at level $p < .05$ after Bonferroni correction. Letters in brackets were significant prior to Bonferroni correction.

A.5. Distribution of additional themes by experimental condition

(Basis: substantive responses)

	A: Open-ended, concurrent	**B: Open-ended, retrospective**	**C: Closed, concurrent**	**D: Closed, retrospective**
	% (n)	% (n)	% (n)	% (n)
P1: Life satisfaction				
Financial situation	28.0% (122)^{c,d}	26.8% (110)^{c,d}	3.3% (18)^{a,b}	2.3% (12)^{a,b}
Corona	21.8% (95)^{c,d}	18.0% (74)^{c,d}	0.9% (5)^{a,b}	1.3% (7)^{a,b}
Living situation	11.2% (49)^{c,d}	12.0% (49)^{c,d}	0.6% (3)^{a,b}	0.6% (3)^{a,b}
Well-being	6.9% (30)^{c,d}	7.3% (30)^{c,d}	0.0% (0)^{a,b}	0.2% (1)^{a,b}
Future and goals	4.8% (21)^{c,d}	7.3% (30)^{c,d}	0.2% (1)^{a,b}	0.4% (2)^{a,b}
Politics, society, environment	5.0% (22)^{c,d}	4.9% (20)^{c,d}	1.1% (6)^{a,b}	1.1% (6)^{a,b}
Private life	7.3% (32)^{c,d}	4.6% (19)^{c,d}	0.0% (0)^{a,b}	0.0% (0)^{a,b}
Lifestyle, standard of living	3.0% (13)$^{(b),c,d}$	5.9% (24)^{c,d}	0.0% (0)^{a,b}	0.0% (0)^{a,b}
Loneliness and stress	2.3% (10)^{c,d}	2.7% (11)^{c,d}	0.0% (0)^{a,b}	0.0% (0)^{a,b}
Other	4.8% (21)^{c,d}	5.6% (23)^{c,d}	0.2% (1)^{a,b}	0.4% (2)^{a,b}
P2: Relationship satisfaction				
Feeling of trust and happiness	15.2% (57)^{c,d}	13.7% (49)^{c,d}	0.0% (0)^{a,b}	0.4% (2)^{a,b}
Harmony and arguments	13.1% (49)^{c,d}	14.8% (53)^{c,d}	0.0% (0)^{a,b}	0.2% (1)^{a,b}
Feeling of love and affection	11.2% (42)^{c,d}	12.6% (45)^{c,d}	0.0% (0)^{a,b}	0.2% (1)^{a,b}
Daily life	10.4% (39)^{c,d}	12.9% (46)^{c,d}	0.2% (1)^{a,b}	0.0% (0)^{a,b}
Treatment and interaction	11.2% (42)$^{(b),c,d}$	6.4% (23)$^{(a),c,d}$	0.4% (2)^{a,b}	0.0% (0)^{a,b}
Communication and humour	9.4% (35)^{c,d}	7.3% (26)^{c,d}	0.0% (0)^{a,b}	0.0% (0)^{a,b}
Character and attitudes	5.3% (20)^{c,d}	5.9% (21)^{c,d}	0.6% (3)^{a,b}	0.6% (3)^{a,b}
Being family and friends	3.7% (14)^{c,d}	1.7% (6)$^{(c),(d)}$	0.0% (0)$^{a,(b)}$	0.0% (0)$^{a,(b)}$
Other	3.5% (13)^{c,d}	5.6% (20)^{c,d}	0.2% (1)^{a,b}	0.7% (4)^{a,b}

(Table A.5 continued)

	A: Open-ended, concurrent	**B: Open-ended, retrospective**	**C: Closed, concurrent**	**D: Closed, retrospective**
	% (n)	% (n)	% (n)	% (n)
P3: Subjective health				
Afflictions	16.2% (63) ^{c,d}	17.2% (65) ^{c,d}	0.0% (0) ^{a,b}	0.0% (0) ^{a,b}
Fitness and sports	14.1% (55) ^{c,d}	11.7% (44) ^{c,d}	0.6% (3) ^{a,b}	0.6% (3) ^{a,b}
Age and independence	7.2% (28) ^{c,d}	10.1% (38) ^{c,d}	0.6% (3) ^{a,b}	0.4% (2) ^{a,b}
Weight, diet, alcohol, smoking	9.0% (35) ^{c,d}	7.7% (29) ^{c,d}	0.2% (1) ^{a,b}	0.7% (4) ^{a,b}
Body & psyche	7.5% (29) ^{c,d}	6.9% (26) ^{c,d}	0.2% (1) ^{a,b}	0.4% (2) ^{a,b}
Well-being	5.9% (23) ^{c,d}	7.2% (27) ^{c,d}	0.0% (0) ^{a,b}	0.2% (1) ^{a,b}
Stress	5.7% (22) ^{c,d}	5.3% (20) ^{c,d}	0.6% (3) ^{a,b}	0.0% (0) ^{a,b}
Physical health	2.8% (11) $^{c,(d)}$	4.0% (15) ^{c,d}	0.2% (1) ^{a,b}	0.4% (2) $^{(a),b}$
Handicap	1.8% (7) $^{(d)}$	2.4% (9) $^{(c),(d)}$	0.7% (4) $^{(b)}$	0.4% (2) $^{(a),(b)}$
Other	4.4% (17) ^{c,d}	5.0% (19) ^{c,d}	0.2% (1) ^{a,b}	0.0% (0) ^{a,b}

Note. Letters in superscript indicate significant differences as per chi-square test of independence at level p < .05 after Bonferroni correction. Letters in brackets were significant prior to Bonferroni correction.

6. STUDY 2: THE IMPACT OF PROBES ON SURVEY QUESTIONS

A version of this chapter has been published as:

Hadler, Patricia (2023). The effects of open-ended probes on closed survey questions in web surveys. *Sociological Methods & Research*, Online First, 1-34. DOI: 10.1177/00491241231176846.

6.1. Introduction

In recent years, open-ended questions have experienced a renaissance (Neuert, Meitinger, Behr, & Schonlau, 2021), particularly in the context of web surveys (Smyth et al., 2009), as the cost and effort of data collection (Gavras & Höhne, 2020; Revilla & Couper, 2019) and coding of responses (Schonlau & Couper, 2016) are much decreased due to technological development. Open-ended narrative questions are considered the "classic open-ended question […], in which respondents are invited to articulate their response using their own words" (Couper, Kennedy, Conrad, & Tourangeau, 2011, p. 67). Probes are a specific type of open-ended narrative question that directly relates to a forgoing closed survey question (Behr et al., 2012b; Schuman, 1966) and can be used to assess the validity and even cross-cultural comparability of survey questions (Meitinger, 2018). They are frequently used at the stage of question development and cognitive pretesting for the purpose of question evaluation, for instance, to examine whether a term in the survey question is understood in the way intended by the researcher or to gain insights on why respondents chose a response option (Collins, 2015; Miller et al., 2014). The implementation of open-ended probes in large-scale surveys is less common, though researchers have repeatedly argued for this to clarify reasons for a response (Schuman, 1966), gain insights on reasons for lack of measurement invariance (Meitinger, 2017), or even to encourage more truthful answers (Couper, 2013; Singer & Couper, 2017).

Next to these described benefits, there are concerns that open-ended probes impact surrounding survey questions. Recent studies that embedded open-ended probes in web surveys indicated an increase in survey break-offs and item nonresponse (Luebker, 2021) as well as slight shifts in response behaviour (Couper, 2013; Fowler & Willis, 2020). If embedding open-ended probes affects the response behaviour to web survey questions, the comparability of survey questions asked with and without open-ended probes may be compromised. This would affect settings such as the one proposed by Schuman (1966), in which probes are asked to a random subsample within a survey, or longitudinal analysis of panel data, if probes are implemented in some, but not all waves. Moreover, the validity of insights gained from web probing responses depends on respondents understanding and answering the survey questions in the same way regardless of whether or not they receive probing questions.

The present article sets out to examine the effects of open-ended probes on web survey questions. The background section begins with an overview of previous studies that examined the effect of open-ended probes on closed survey questions and points to current research gaps. Next, the differences between open-ended and closed questions in terms of burden, cognitive processing and response behaviour are summarized, and the notion of measurement reactivity in surveys is introduced. From this, the research questions and hypotheses are derived. A between-subject experiment is reported which assessed the effects of embedding open-ended probes on the processing of and response to closed web survey questions. The experiment examined survey break-off, backtracking, and answer changes to previous survey questions, as well as response times, nonresponse, and response behaviour to successive survey questions. Finally, the benefits and potential adverse effects of embedding open-ended probes into web surveys are discussed.

6.2. Background

6.2.1. Previous research on the impact of open-ended probes on survey response

In the realm of web surveys, three studies have examined the impact of open-ended probes on closed survey questions. Luebker (2021) examined the effect of embedding an open-ended probe on survey break-off and item nonresponse to a closed opinion question. He found that a probe displayed on the same survey page as the question it pertained to increased survey break-off by 0.6 and item nonresponse by more than 25 percentage points. When using a paging design—that is displaying the probe on a separate survey page—there was a stronger impact on survey break-off of 1.4 percentage points, but no effect on item nonresponse as compared to inserting no probe. It must be noted that the probe in this experiment more resembled an open-ended text field at the end of a closed survey question than a typical open-ended narrative probe and was worded in a strongly nonmandatory manner ("If you like, you can add some bullet points to your response.").

Couper (2013) reported the results of two experiments that inserted open-ended probes into a ten-item scale on attitudes towards immigrants in a probability panel in the Netherlands. In the first experiment, respondents were presented with one mandatory open-ended probe after each item using a paging design. In the second experiment, respondents were presented with an optional open-ended probe on the same screen as the respective closed survey item. In both experiments, there was a small but significant difference in the overall means of the item battery between the experimental condition with probes and the control group, with respondents reporting lower levels of prejudice in the condition with probes. In the experiment using the paging design, this effect occurred as of the second-shown survey item.

Finally, a study by Fowler and Willis (2020) compared responses to survey questions depending on probe placement. Respondents answered nine closed items on perceptions of neighbourhood walkability, such as the presence of sidewalks, trails, or paths. On one condition, they received four open-ended probes on the survey page directly after the item battery. In the other condition, respondents were presented with the probes retrospectively at the end of the survey, that is, with several unrelated survey questions in between. Results showed a small but significant effect of probe placement on the mean walkability score, with respondents who received the probes directly after the survey questions reporting slightly enhanced perceptions of walkability. It must be noted that the study was not a strictly randomized experiment, as the condition that included probes directly after the survey items was fielded three weeks before the condition with retrospective probes, and the sample was not representative of the U.S. population in terms of demographics.

In sum, previous studies lend support for the notion that open-ended probes impact whether a respondent continues the survey, answers survey questions, and how they answer them. However, the studies also raise many questions. Regarding survey break-off and item nonresponse, the effect sizes found by Luebker (2021) merit further examination. The study only examined one survey question, and the probe was rather atypical. Effects may vary across both probe and question types. Regarding response behaviour to the survey questions, Couper (2013) found that using a paging design influenced response behaviour to subsequent items (i.e., as of the second-shown item), whereas Fowler and Willis (2020) found an effect on their item battery despite presenting their probes on a separate page after the survey questions. A possible explanation for this effect could be that respondents in Fowler and Willis' study backtracked to the previous survey page and changed their answers; however, the study provides no details on this. Most importantly, the reasons for the shift in the overall means found in both studies remain unclear. Couper (2013) had assumed that open-ended text fields in which respondents could justify their responses would reduce the threat of sensitive questions and lead to an increase in socially undesirable answers; however, the shift in response behaviour indicated the opposite effect. Moreover, in both studies, the effects on the overall means were rather small. Possibly, effects on response behaviour can be better examined using other measures, such as indicators of response quality or response styles. However, there is currently no framework to predict such effects.

6.2.2. Cognitive processing of open-ended and closed questions

The following section draws on literature on open-ended probes in the context of cognitive interviewing and web probing, as well as open-ended narrative questions in general. Other types of open-ended (such as numeric) questions or probes with closed response options, also known as targeted embedded probes (Scanlon, 2019, 2020), are not considered.

The process of survey response optimally consists of several cognitive steps (Tourangeau et al., 2000). Respondents must interpret the pragmatic meaning of a survey question. They embark on an information retrieval process, which is truncated when respondents have gathered enough information to form a judgment of sufficient certainty. The relevance of the accessible information is assessed, and an internal judgment is formed. This is then adjusted to the response format of the survey question.

For closed questions, the available response options may contribute to construing the meaning of a question (Schwarz et al., 1988) and impact the perceived relevance of the retrieved information. For open-ended questions, neither question interpretation nor the assessment of which accessible information is relevant to form a judgment is guided—and potentially limited—by predefined response options. However, open response formats also bear the risk that respondents deem aspects irrelevant if they consider them self-evident, or that information retrieval is truncated before relevant information is retrieved (Tourangeau et al., 2014).

The differences in these processes impact whether and how respondents answer open-ended and closed questions. In general, the respondent tasks associated with open-ended questions are considered more demanding and burdensome (Krosnick, 1999; Tourangeau & Rasinski, 1988). In line with this, a higher number of open-ended questions in a survey is associated with an increased likelihood of survey break-off (Galesic, 2006), and inserting multiple open-ended questions on one page has a particularly strong effect (Peytchev, 2009). The study by Luebker (2021) confirmed the negative impact on survey break-off for open-ended probes, in particular when the probe is presented on a separate survey page. Moreover, open-ended questions result in higher levels of item nonresponse than corresponding closed questions (Reja et al., 2003; Zuell et al., 2015), particularly among lower educated respondents (Andrews, 2005; Miller & Lambert, 2014; Schmidt et al., 2020; Scholz & Zuell, 2012; Zuell & Scholz, 2015). These findings have been confirmed for open-ended as compared to closed probes (Neuert, Meitinger, & Behr, 2021).

Differences can also be found in response distributions between open-ended and closed survey questions (Reja et al., 2003) and probes (Neuert, Meitinger, & Behr, 2021). Responses not included in a closed format are unlikely to be named by respondents, even when an open-ended “other” field is included. On the other hand, any given opinion, theme, or topic is less likely to be volunteered in an open response format than when it must simply be “recognized” in a closed question (Bradburn, 1983).

Open-ended probes are more directed than other open-ended questions as they directly pertain to the preceding closed survey question (Foddy, 1998; Neuert, Meitinger, Behr et al., 2021; Silber et al., 2020). In the context of web probing, three types of probes are mainly employed. Comprehension probes ask about the understanding of a term used in the survey question. In category selection probes respondents are requested to explain why they chose their response option. Specific probes encourage respondents to provide additional information on a particular detail of the item (Behr et al., 2012b; Behr, Braun et al., 2012; Meitinger et al., 2018). These probe types ask respondents to focus on different aspects of their survey responses. Due to the differences in the response process between open-ended and closed questions, probes may lead respondents to consider additional or different aspects of a question than they did while answering it, or respondents may come to a different evaluation as to which retrieved information should be relevant to their judgment. Moreover, simply the process of repeated thinking about survey questions may impact how a respondent answers them. This could lead to an interaction of probing and survey questions. The following section describes the impact that questions within a survey can have on each other, known as measurement reactivity, and applies it to probing questions.

6.2.3. Measurement reactivity in surveys

The notion that examining a phenomenon can alter the phenomenon itself is discussed in many areas of research, from physics to behavioural psychology. In survey research, the notion of measurement reactivity was examined in a series of experiments using personality measures. Knowles et al. (1992) argued that thinking about questions has consequences for question construal and that increased reflection on a topic makes a certain interpretation more salient, leading to a polarization of judgment. They postulated that later items within a measure (or items in a repeated measurement) show more extreme, but also more reliable and consistent responses. To examine this, the order of multi-item measures was randomized (Knowles, 1988), with later items showing higher reliability and more extreme answers. Importantly, there was generally no visible effect on the mean value of these items. The studies demonstrated that increased reflection about survey questions influences both cognitive processing and response to survey items and that these effects must not (necessarily) be visible by a simple comparison of means.

Whether respondents' verbalized reflection on survey questions causes reactivity has been subject to debate since the dawn of cognitive testing. The early standard of verbal protocols required respondents to think aloud while answering a survey question (Ericsson & Simon, 1980, 1993). This was criticized by researchers as potentially increasing the effort required to create a response (Willis, 1994, 2005), especially after an experimental study demonstrated that think-aloud protocols impact task accuracy and response times for some tasks (Russo et al., 1989). The debate gave rise to the use of probing questions, which are administered after the respondent has completed the survey question. Beatty and Willis (2007) argued that using probes in cognitive interviews may be less likely to cause reactivity than employing the think-aloud technique. However, other researchers argued that probes may likewise lead to invalid or reactive reports

(Conrad et al., 1999) by interfering with the natural flow of the survey interview (Beatty, 2004), and a recent meta-analysis supported the notion that directive probing can impact task accuracy (Fox et al., 2011). A further study showed some indication of increased respondent motivation through verbal probing, but remained inconclusive (Sudman, Bradburn, & Schwarz, 1996).

6.3. Research questions and hypotheses

This study aims to enhance our understanding of the impact of open-ended probes on survey responses. I differentiate between effects on the survey in terms of survey completion, effects on the questions being probed, and effects on subsequent questions.

RQ1: Does embedding open-ended probes into web surveys impact survey break-off?

RQ2: Does embedding open-ended probes into web surveys impact the survey questions the probes pertain to?

RQ3: Does embedding open-ended probes into web surveys impact subsequent survey questions?

Based on the findings from previous studies and literature on open-ended questions and measurement reactivity, I put forward several hypotheses. The strongest possible adverse effect of an open-ended question occurs if a respondent chooses to discontinue the survey. The sum of past research indicates that adding open-ended probes to a web survey results in higher levels of survey break-off (Galesic, 2006; Luebker, 2021; Peytchev, 2009).

H1: Embedding open-ended probes into web surveys increases survey break-off.

Embedding open-ended probes may impact the survey questions they relate to, either if probes are presented alongside the survey question on the same page (as in some of the experimental conditions in (Couper, 2013; Luebker, 2021), or if respondents have the possibility to return to previous questions in a paging design (which would explain the effects found by Fowler & Willis, 2020). A probe may cause respondents to reconsider their interpretation of a survey question, access other information, or include other information in their judgment. Previous research has indicated that reverse question order effects may arise when respondents have the possibility to return to previous questions (Sudman et al., 1996), meaning that subsequent questions can influence responses to previous ones (Bishop, Hippler, Schwarz, & Strack, 1988; Schwarz & Hippler, 1995; Schwarz, Strack, Hippler, & Bishop, 1991). Therefore, I hypothesize that embedding open-ended probes leads to an increase in backtracking and changing one's answer to previous survey questions:

H2a: Embedding open-ended probes into web surveys increases backtracking to previous survey questions.

H2b: Embedding open-ended probes into web surveys increases answer changes to previous survey questions.

Next to the effects on the survey questions they relate to, open-ended probes may impact how respondents process and answer subsequent questions. The following hypotheses rest on the assumption that probes cause respondents to reflect on their previous survey responses, and that respondents process survey questions more deeply when they are expecting these questions to be followed by probes. Knowles (1988) demonstrated that increased thinking about questions leads to judgment polarization and more consistent responses.

Regarding cognitive effort, response times are considered "one of the most important means for investigating hypotheses about mental processing" (Yan & Tourangeau, 2008, p. 51). Findings from think-aloud and verbal probing (Fox et al., 2011; Russo et al., 1989) indicate that response times increase in interviewer-administered settings for some questions. Therefore, I hypothesize that response times to closed survey questions increase when open-ended probes are embedded:

H3: Embedding open-ended probes into web surveys increases response times to subsequent survey questions.

Unfortunately, previous studies did not report on nonresponse (Couper, 2013; Fowler & Willis, 2020) or only examined one question (Luebker, 2021). However, if embedding open-ended probes causes respondents to reflect on survey questions more deeply and this leads to judgment polarization, it can be assumed that nonresponse decreases for subsequent items:

H4: Embedding open-ended probes into web surveys decreases nonresponse for subsequent survey questions.

Previous research has found effects of embedding open-ended probes on the mean sum score of multi-item measures (Couper, 2013; Fowler & Willis, 2020), but the effect size was small and the direction could not be predicted. Knowles et al. (1992) argued that increased thinking about answers impacts response behaviour but that this must not necessarily be visible in the form of a mean shift. Rather, increased thinking about questions leads to judgment polarization and more consistent responses. Therefore, I hypothesize that embedding open-ended probes does not (consistently) impact means. Instead, differences become visible in the form of increased extreme responding and non-differentiation:

H5a: Embedding open-ended probes into web surveys does not impact mean scores for subsequent survey questions.

H5b: Embedding open-ended probes into web surveys increases extreme responding for subsequent survey questions.

H5c: Embedding open-ended probes into web surveys increases non-differentiation for subsequent survey questions.

6.4. Method

An experiment was designed with the aim of comparing closed survey questions that were either accompanied by open-ended probes or not. Respondents received six survey pages with closed attitude questions. A between-subject design was used, in which respondents were randomly assigned to experimental condition (A) which embedded open-ended probes between the survey pages with closed questions, or condition (B) which contained only the closed questions.

6.4.1. Survey and probing questions

The closed survey questions presented a mix of single- and multi-item measures using common constructs in social science research, such as political attitudes, personality, and well-being. Measures that have been accompanied by open-ended probes in other studies were chosen when possible. Multi-item measures were presented in a grid format on one survey page. The exact wording of the closed survey questions and the open-ended probes from condition (A) can be found in Table A.1 of the Appendix.

The first closed survey question was a single-item measure of left-right orientation (Q1), which is considered a "central element of political science research" (Zuell & Scholz, 2015, p. 28). Left-right orientation is implemented in several general population surveys, such as the German General Social Survey (GESIS, 2020) or the German Longitudinal Election Study (GLES, 2019). The question has repeatedly been complemented by open-ended probes to gain additional insights into the meaning of left and right (Bauer, Barbera, Ackermann, & Venetz, 2017; Fuchs & Klingemann, 1989, 1990; Scholz & Zuell, 2012; Zuell & Scholz, 2015). It was succeeded by another political construct, the two-item short scale on political cynicism (Q2) (Aichholzer & Kritzinger, 2016), which captures respondents' general trust in politicians' honesty. It was developed for the Austrian National Election Study (Kritzinger et al., 2014) and has been implemented by other researchers since (Prochazka, 2020). The third question was the six-item short scale for the Gamma factor of social desirability (Q3) (Kemper, Beierlein, Bensch, Kovaleva, & Rammstedt, 2014). Social desirability responding is the tendency to give overly positive self-descriptions (Paulhus, 2002). The construct's Gamma factor is implemented into questionnaires to control whether self-reports may be biased by social desirability responding (Nießen, Partsch, Kemper, & Rammstedt, 2019). Survey research has indicated reactivity in personality measures (Knowles, 1988; Knowles et al., 1992). General life satisfaction (Q4) and relationship satisfaction (Q5) (Beierlein et al., 2014; Schwarz, Strack, & Mai, 1991) were implemented as further single-item measures. They have been followed by open-ended probes in the past to analyse determinants of self-reported satisfaction (Edwards & Lopez, 2006). Furthermore, past research has repeatedly demonstrated that the communicative context of these questions impacts how the items are answered and how strongly they relate to one another (Schuman & Presser, 1981; Smith, 1982). The measure of intergenerational social support (Q6) consists of six items on family support (Gerlitz, 2014; Legewie, Gerlitz, Mühleck, Scheller, & Schrenker, 2007). All questions included the non-substantive response option “I don’t want to answer” except the personality measure (Q3).

In experimental condition (A), each closed question was followed by one open-ended probe. For the multi-item inventories (Q2, Q3, and Q6), respondents were randomly presented with a comprehension or category selection probe. The single-item measures on life and relationship satisfaction (Q4 and Q5) were each followed by a specific probe. The question on left-right orientation (Q1) was followed by two probes on the understanding of the terms “left” and “right” as in previous studies (Zuell & Scholz, 2015). To keep the survey setting identical across conditions and to be able to attribute survey break-offs to the probing situation, the probes were not announced on the welcome page of the survey. Instead, the first probe was introduced with the words “We would like to receive more information on the previous question.” Probes were presented using a paging design, with the survey question being repeated on the page with the probe. In addition, the selected survey response was repeated for category selection and specific probes.

6.4.2. Web survey

An online survey was carried out with a nonprobability sample between November 20th and December 2nd, 2020, with the panel provider Respondi AG. The sample included quotas to depict the German online population in terms of gender (male, female),[13] age (18-29, 30-39, 40-49, 50-59, and 60 or more years), education (low, medium, and high) and region (former East and West Germany). Respondents were randomly assigned to experimental condition (A) or (B). There were no significant differences regarding demographics or devices used between experimental groups (see Table A.2 in the

[13] Respondents could also choose the nonbinary category “divers”; this was, however, not subjected to quotas.

Appendix for the sample composition). The web survey included several experiments, which were randomized independently of each other. The reported study was placed towards the beginning of the survey after the screening and quota questions and three short scales. No open-ended questions were implemented before the experiment.

The Universal Client-Side Paradata script by Kaczmirek and Neubarth (2007) was implemented to ensure a more exact measure of response latency (Yan & Tourangeau, 2008) and collect questionnaire navigation data (Callegaro et al., 2015). The script records response behaviour sequentially so that the resulting string variables enable coding backtracking to previous survey pages and answer changes to items. In accordance with both legal and ethical research standards (ADM et al., 2021; Kunz et al., 2020), respondents were informed about the collection and use of client-side paradata on the welcome page of the survey.

Of the 9,731 panelists invited to participate in the survey, 2,441 started the survey and of those, 241 broke off before completing it (before, during, or after the reported experiment), resulting in 2,200 completed questionnaires. This leads to a participation rate of 22.6% (American Association for Public Opinion Research [AAPOR], 2016) and a break-off rate of 9.9% (Callegaro & DiSogra, 2008). Respondents received €1.50 for survey completion. About a quarter (27.1%; n=597) of respondents filled out the survey using a mobile device. Average survey completion was 21.0 min (median: 17.3; n=1,096) for respondents in condition (A) and 15.9 min (median: 12.4; n=1,046) for respondents in condition (B).

6.4.3. Probe response quality and content

Prior to examining the survey responses, probe response quality and content were analysed. Low probe response quality may point to poorly designed probing questions. Probe response content was examined to gain insights into the respondents' cognitive process of survey response and the quality of the survey questions. To ascertain probe response quality, the share of non-substantive responses was determined. Responses were coded as non-substantive when respondents left the probe empty, entered random characters, typed a "don't know" answer, an explicit refusal, or other non-intelligible or non-codable content (Behr, Braun et al., 2012). Between 12.2% and 21.6% of respondents in condition (A) gave non-substantive responses to the probing questions (see Table 6.1), which coincides with previous web probing studies (Behr et al., 2017; Meitinger & Behr, 2016). To examine probe response content, the substantive probe responses were subjected to cognitive coding (Willis, 2015a) to determine whether the probe responses hinted at issues in respondents' cognitive process of survey response. This approach is also known as an error perspective, as it gives insights into possible reasons behind measurement errors (Meitinger & Behr, 2016). Errors or problems may occur at the stages of question comprehension, information retrieval, judgment, or response formatting (Willis, Schechter et al., 1999). For all but one question, error codes were detected for under 5% of respondents (see Table 6.1). Reported problems included misinterpreting words central to the question (i.e., the terms "left" or "right" in Q1) or mismatches between survey and probe responses (i.e., a low score of political cynicism in Q2, but probe responses indicating very low trust in politicians). Question Q5 on relationship satisfaction showed an unusually high share of error coding, with 25.5% of responses pointing to difficulties with judgment and/or response formatting. Almost all of these respondents reported that they were missing an answer to indicate that they were currently not in a relationship. A complete list of reported problems is available from the author on request.

Table 6.1. Probe response quality and content

	Non-substantive response	**Error detected**	**Error related to…**	
			Question comprehension / Information recall	**Judgment / Response formatting**
Q1*	21.6% (237)	4.7% (51)	3.7% (41)	1.0% (11)
Q2	13.0% (143)	4.7% (51)	2.1% (23)	2.6% (29)
Q3	13.8% (151)	1.9% (21)	0.5% (6)	1.4% (15)
Q4	12.2% (134)	0.9% (10)	0.2% (2)	0.7% (8)
Q5	15.5% (170)	26.3% (288)	0.7% (8)	25.5% (280)
Q6	19.3% (212)	3.4% (37)	2.2% (24)	1.2% (13)

*For Q1, responses were coded as non-substantive when both probes contained non-substantive content; an error was coded when at least one probe response contained the error code.

Note. *Based on condition A, N = 1,096*

6.4.4. Dependent measures and data analysis

All dependent measures were compared across the two experimental conditions (1=condition (A) open-ended probes; 0=condition (B) without probes).

Survey break-off, backtracking, and answer changes. Break-offs that occurred during the reported experiment, that is on the pages containing the closed survey questions (Q1-Q6) or the open-ended probes (P1a/b-P6) were included in the analysis. Backtracking was recorded for respondents when the client-side paradata string recorded more than one page visit. Multiple page visits were coded into a binary variable for each respondent on page level (1 = backtracking to survey page; 0 = no backtracking to survey page) and aggregated across all six closed survey questions (1 = backtracking to at least one survey page; 0 = no backtracking to any survey pages). Likewise, the prevalence of answer changes after backtracking was coded on page level (Heerwegh, 2003, 2011) (1 = answer change after backtracking; 0 = no answer change/no backtracking) and aggregated across all six closed survey questions (1 = answer change after backtracking on at least one survey page; 0 = no answer change after backtracking for any closed survey questions).

Chi-square tests of independence were used to examine whether the share of survey break-offs, backtracking, and answer changes after backtracking differed between conditions. For backtracking and subsequent answer changes, analyses were additionally carried out on page level, with Bonferroni adjusted alpha levels for multiple comparisons.

Response times. Client-side response times were captured for each survey page. Response time data is positively skewed and subject to outliers, which makes decisions about outlier definition, handling, and potential transformation of response times prior to analysis of utmost importance (Kunz & Hadler, 2020). A variety of response time outlier definitions exist (Matjašič et al., 2018). Detecting outliers based on the mean and standard deviation remains a common procedure (Yan & Olson, 2013), but has been criticized, as the mean value is in turn influenced by outliers. Researchers have increasingly recommended using median-based outlier definitions (Höhne & Schlosser, 2018). Employing 2.5 times the median absolute deviation is considered the most robust outlier threshold (Leys, Ley, Klein, Bernard, & Licata, 2013), and applied in the present study. This method led to between 9% and 12% of response times being identified as outliers.

Response time outliers were omitted from response time analysis, as were all instances in which a survey page was visited more than once (backtracking).

To test the third hypothesis, a multivariate analysis of covariance (MANCOVA) was applied to the adjusted response times as of the second survey question (Q2-Q6), with the experimental condition as the main predictor. The model included age, education, and device used (PC/laptop or mobile) as covariates, as it was not possible to include baseline reading speed, which otherwise accounts for much of the variance within response times (Couper & Kreuter, 2013; Lenzner et al., 2010; Yan & Tourangeau, 2008). Because outlier exclusion as described above leads to a sizeable decrease in available cases for a MANCOVA, separate analyses of covariances (ANCOVAs) were run for each survey page (thus, only excluding outliers page-wise).

The robustness of response time analyses should be tested by applying different outlier detection methods (Revilla & Couper, 2018). The alternative approach excluded observations beyond the upper and lower one percentile from analysis (Yan & Tourangeau, 2008). A MANCOVA and separate ANCOVAs were run using the same predictors. All analyses led to the same results.

6.4.5. Response behaviour

Survey response. T-tests were used to test for differences between conditions. For the single-item measures (Q4 and Q5), the scale values coincide with the items' raw values. For the multi-item measures, scoring was carried out according to the instruments' documentation. For the two-item measure on political cynicism (Q2), the second item was recoded so that the scale direction was the same across items, with higher values indicating higher cynicism. The score is the sum of both items divided by the number of items (Aichholzer & Kritzinger, 2016). For the social desirability responding measure (Q3), the two factors exaggerating positive qualities (PQ+) and minimizing negative qualities (NQ−) were calculated separately using the sum score, again dividing by the number of items (Kemper et al., 2014). For the intergenerational social support inventory (Q6), the fourth item was recoded so that the scale direction remained the same across items, and principal components factor analysis was conducted on the six items with oblique rotation (Direct Oblimin). The Kaiser-Meyer-Olkin measure of sampling adequacy was 0.65, and Bartlett's test of sphericity reached statistical significance

($\chi^2(15) = 1{,}947.06$, $p < .001$). The analysis identified two factors with an Eigenvalue greater than 1, explaining 58.84% of the variance. All items loaded above .60. The two factors could be attributed to support provided by older to younger generations (F1) and support provided by younger to older generations (F2).

Item nonresponse. Item nonresponse can occur in the form of (1) skipping an item entirely, that is, when respondents produce missing data by leaving an item blank, or (2) by choosing a non-substantive response option, such as “I don’t want to answer” or “I don’t know” (Cornesse & Blom, 2020). Skipping items is the more precise indicator of satisficing and low response quality, as choosing a non-substantive response option at the least requires reading and choosing the respective response option (Schuman & Presser, 1981). Because these types of nonresponse may occur for different reasons, they may be impacted differently by embedding open-ended probes and are analysed separately.

Chi-square tests of independence were used to examine whether the frequency of skipping items or choosing non-substantive response options differed between conditions. Analyses were additionally carried out on page level (for Q2-Q6), with Bonferroni-adjusted alpha levels for multiple comparisons. For multi-item inventories (Q2, Q3, and Q6), nonresponse was aggregated across all items. The item battery on social desirability responding (Q3) did not include a non-substantive response option and is not included in this analysis.

Non-differentiation. Non-differentiation, also referred to as straightlining, was examined for the two multi-item batteries Q3 and Q6. A dichotomous and a metric measure of straightlining were calculated. The dichotomous measure indicated whether a respondent chose the identical response option for all items within one battery. The second measure “mean root of pairs” was chosen to “capture variations in the choice of answers in a battery” (Kim et al., 2019, p. 227). It is calculated by producing a temporary index by computing the mean of the root of the absolute differences between all pairs of items in a battery and then rescaling the temporary index to range from 0, indicating least straightlining, to 1, indicating most straightlining (Chang & Krosnick, 2009). To examine whether non-differentiation differed between conditions, chi-square tests of independence were used for the absolute and T-tests for the metric measure of non-differentiation.

Extreme responding. Extreme responding was defined as choosing either endpoint of a response scale (apart from non-substantive response options). Chi-square tests of independence were carried out on item level with Bonferroni-adjusted alpha levels for multiple comparisons.

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

6.5. Results

6.5.1. Effect on survey break-off

The first hypothesis postulated that embedding open-ended probes into a web survey increases survey break-off. Seventy-nine respondents dropped out of the survey during the reported experiment. Most of these break-offs (82.3%; *n* = 65) occurred in condition (A) with open ended probes. This difference was significant (*n* = 2,279; $\chi^2(1)=32.153$; $p < .001$); therefore hypothesis 1 can be confirmed.

More than half of the break-offs in condition (A) (*n* = 44) occurred during the first two survey questions (Q1 and Q2) or their respective probes (P1a, P1b, and P2). Respondents who broke off the survey during the reported experiment were significantly more likely to have lower education (low/medium: *n* = 59 vs. high: *n* = 20; $\chi^2(1)=7.567$; $p = .006$) and be women (women: *n* = 55 vs. men: *n* = 24; $\chi^2(1)=12.914$; $p < .001$), consistent with previous research (Roßmann et al., 2015) that respondents who break off surveys systematically differ from those who complete them.

6.5.2. Effect on preceding survey questions

The second hypothesis stated that embedding open-ended probes into a web survey increases backtracking (H2a) and subsequent answer changes (H2b) to previous survey questions. In total, 6.1% (*n* = 134) of respondents backtracked to a previous survey page with closed questions. Of these, three quarters (76.1%; *n* = 102) were assigned to condition (A) with open-ended probes. A chi-square test of independence confirmed the significant difference between conditions ($\chi^2(1) =39.483$; $p < .001$). Backtracking to the first-shown question (Q1) occurred most often (*n* = 46). Analysis of question level showed the same tendency for all six questions (see Table 6.2). Differences between conditions were significant for Q1, Q3, and Q4 based on Bonferroni-adjusted alpha levels of .0083 per test (.05/6).

Answer changes after backtracking were carried out by 2.0% of respondents (*n* = 45). Of these, 71.1% (*n* = 32) were respondents in condition (A) with open-ended probes. A chi-square test of independence confirmed that the difference between conditions was significant ($\chi^2(1) = 8.332$; *p* = .004), though there were no significant differences in question level when using the Bonferroni adjusted alpha levels (see Table 6.2). It should be noted that across both conditions, a similar share of respondents who returned to a previous survey page decided to change their answers. Thus, while embedding open-ended probes increases backtracking and subsequently answer changes, it does not increase the likelihood of backtrackers to change their survey response. In summary, hypothesis 2 can be confirmed.

Table 6.2. Backtracking and answer changes on question level

	Valid *n*	**Condition (A): open-ended probes**	**Condition (B): without probes**	**Chi-square test of independence**
		% (*n*)	% (*n*)	$\chi^2_{(df)}$; *p*
Backtracking to previous survey question				
Q1	2,197	3.3% (36)	0.9% (10)	$\chi^2_{(1)} = 15.180$; *p* < .001
Q2	2,196	1.4% (15)	0.6% (7)	$\chi^2_{(1)} = 2.998$; *p* = .083
Q3	2,196	2.0% (22)	0.5% (5)	$\chi^2_{(1)} = 10.963$; *p* = .001
Q4	2,196	1.3% (14)	0.1% (1)	$\chi^2_{(1)} = 11.440$; *p* = .001
Q5	2,196	1.3% (14)	0.5% (6)	$\chi^2_{(1)} = 3.289$; *p* = .070
Q6	2,196	1.4% (15)	0.6% (7)	$\chi^2_{(1)} = 2.998$; *p* = .083
Answer change after backtracking				
Q1	2,197	0.7% (8)	0.4% (4)	$\chi^2_{(1)} = 1.370$; *p* = .242
Q2	2,196	0.5% (5)	0.4% (4)	$\chi^2_{(1)} = 0.119$; *p* = .730
Q3	2,196	0.5% (5)	0.1% (1)	$\chi^2_{(1)} = 2.703$; *p* = .100
Q4	2,196	0.4% (4)	0.1% (1)	$\chi^2_{(1)} = 1.826$; *p* = .177
Q5	2,196	0.8% (9)	0.1% (1)	$\chi^2_{(1)} = 6.488$; *p* = .011
Q6	2,196	0.5% (5)	0.2% (2)	$\chi^2_{(1)} = 1.312$; *p* = .252

6.5.3. Effect on subsequent survey questions

Response times. The third hypothesis stated that embedding open-ended probes into a web survey increases cognitive effort and thus response time invested in subsequent questions. To examine this, the response times to Q2-Q6 between conditions were examined using MANCOVA. After outlier exclusion, 1,509 cases remained for analysis (condition A: $n = 737$; condition B: $n = 772$). There was a significant main effect of condition on overall response time (Wilks-Lambda = 0.98; $F_{(5,1500)} = 6.01$; $p < .001$; η2=.02). Age and device used were significant covariates, whereas education was not.2[14] Table 6.3 depicts the mean response times and separate ANCOVAs run for each survey question. Results showed that response times differed significantly for only one of the examined questions: respondents who were presented with open-ended probes spent significantly longer on average (7.4s) to respond to the question on relationship satisfaction (Q5) than respondents in condition (B) (6.7s). Thus, based on the result of the MANCOVA, hypothesis 3 can be confirmed; however, the effect only seems to apply in specific situations.

Table 6.3. Survey question response times, ANCOVAs

	Q2	**Q3**	**Q4**	**Q5**	**Q6**
N	1,999	1,977	1,954	1,971	2,012
Mean response time s (std dev)					
Condition (A)	16.7 (7.5)	40.9 (17.9)	8.0 (3.4)	7.4 (3.3)	45.8 (21.5)
Condition (B)	16.5 (7.5)	40.5 (18.1)	8.1 (3.7)	6.7 (3.2)	46.9 (23.0)
F	1.12	1.73	0.15	30.68	0.45
p	0.29	0.19	0.70	< 0.001	0.50
partial η2	0.00	0.00	0.00	0.02	0.00

Nonresponse. The fourth hypothesis postulated that nonresponse is lower in the condition with open-ended probes. Table 6.4 shows the occurrence of skipping items and choosing non-substantive response options on question level. Across Q2-Q6, 8% ($n = 175$) of respondents skipped at least one item. A chi-square test of independence showed no significant difference between conditions ($\chi^2(1) = 0.251$; $p = .616$) across the

[14] MANCOVA results with respect to the covariates (1) age: Wilks-Lambda = 0.87; $F_{(5,1500)} = 46.59$; $p < .001$; η2 = .13; (2) device used: Wilks-Lambda = 0.92; $F_{(5,1500)} = 25.76$; $p < .001$; η2=.08; (3) education: Wilks-Lambda = 0.996; $F_{(5,1500)} = 1.18$; $p = .319$ n.s.

five questions. Question level tests using Bonferroni adjusted alpha levels of .01 (.05/5) confirmed this, though based on the global alpha level, more respondents left the question on relationship satisfaction (Q5) unanswered in condition (B) without probes.

Table 6.4. Nonresponse

	Condition (A): open-ended probes	**Condition (B): without probes**	**Chi-square test of independence**
	% (*n*)	% (*n*)	$\chi^2_{(df)}$; *p*
Item skipping			
Q2	0.7% (8)	1.3% (14)	$\chi^2_{(1)}$ = 1.61; *p* = .205
Q3	4.3% (47)	3.9% (43)	$\chi^2_{(1)}$ = 0.22; *p* = .641
Q4	0.1% (1)	0.1% (1)	$\chi^2_{(1)}$ = 0.00; *p* = .996
Q5	0.4% (4)	1.3% (14)	$\chi^2_{(1)}$ = 5.53; *p* = .019
Q6	2.6% (28)	2.5% (28)	$\chi^2_{(1)}$ = 0.00; *p* = .978
Non-substantive response			
Q2	4.7% (52)	5.2% (57)	$\chi^2_{(1)}$ = 0.21; *p* = .651
Q4	2.1% (23)	0.8% (9)	$\chi^2_{(1)}$ = 6.32; *p* = .012
Q5	16.0% (175)	13.3% (147)	$\chi^2_{(1)}$ = 3.10; *p* = .078
Q6	6.8% (75)	6.3% (69)	$\chi^2_{(1)}$ = 0.32; *p* = .574

Respondents had the option to choose a non-substantive answer (“I don’t want to answer”) for all questions except Q3. In total, 22.7% (*n* = 499) of respondents chose a non-substantive response to at least one item. A chi-square test of independence showed a significant difference between conditions (χ^2(1) = 3.905; *p* = .048). However, contrary to the hypothesis, more respondents in condition (A) with open-ended probes chose a non-substantive response (24.5%; *n* = 268) than in condition (B) (20.9%; *n* = 231). This difference remained significant on the question level for the question on general life satisfaction (Q4) based on the Bonferroni adjusted alpha level of .0125 (.05/4). Thus, hypothesis 4 could not be confirmed.

Mean scores. In line with hypothesis 5a, there were no significant differences between means for any of the single-item or multi-item measures (see Table 6.5).

Table 6.5. Mean scores

	Range	**Condition (A): open-ended probes** **mean (std)**	**Condition (B): without probes** **mean (std)**	**$T_{(df)}$; *p***
Q1	1 (left) to 10 (right)	5.30 (1.78)	5.26 (1.86)	$T_{(1918)}$ = .52; *p* = .600
Q2	1 (low) to 5 (high)	3.63 (.86)	3.66 (.88)	$T_{(2067)}$ = -.63; *p* = .528
Q3 PQ+	1 (low) to 5 (high)	3.58 (.70)	3.53 (.77)	$T_{(2126)}$ = 1.62; *p* = .106
Q3 NQ-	1 (low) to 5 (high)	2.04 (.82)	2.03 (.81)	$T_{(2154)}$ = .38; *p* = .707
Q4	1 (low) to 11 (high)	7.19 (2.19)	7.03 (2.35)	$T_{(2163)}$ = 1.58; *p* = .114
Q5	1 (low) to 11 (high)	7.90 (3.19)	7.69 (2.91)	$T_{(1830)}$ = 1.46; *p* = .143
Q6: F1	-2.79 (disagree) to 3.15 (agree)	-.02 (1.01)	.02 (.99)	$T_{(1997)}$ = -.83; *p* = .406
Q6: F2	-2.55 (disagree) to 2.87 (agree)	-.01 (1.00)	.01 (1.00)	$T_{(1997)}$ = -.50; *p* = .616

Extreme responding. According to Hypothesis 5b, more extreme responding should occur in condition (A) with open-ended probes. Chi-square tests were conducted for each of the eight probed items of Q2-Q6, using Bonferroni adjusted alpha levels of .00625 (.05/8). Table 6.6 shows the share of extreme responding across conditions on item level. Extreme responding was more likely for the single-item measure of relationship satisfaction (Q5). In line with the hypothesis, significantly more respondents reported that they were extremely satisfied or unsatisfied with their current relationship in condition (A) that included open-ended probes (37.4%, *n* = 343) than when no probes were embedded (25.7%; *n* = 242) (χ^2(1)=29.73; *p* < .001). There were no further significant differences between conditions. Thus, similar to the findings regarding response times, hypothesis 5b can be confirmed for one question only, that being the question on relationship satisfaction.

Non-differentiation. Hypothesis 5c assumed higher levels of non-differentiation among respondents in condition (A) with open-ended probes. This was examined for the two multi-item batteries Q3 and Q6. However, neither the absolute nor the metric measure using the mean root of pairs showed significant differences between conditions (see Table 6.7). Therefore, an impact on non-differentiation cannot be confirmed.

Table 6.6. Extreme responding (probed items only)

Question	**Valid *n***	**Condition (A): open-ended probes**	**Condition (B): without probes**	**Chi-square test of independence**
		% (*n*)	% (*n*)	$\chi^2_{(df)}$; *p*
Q2 (item 1)	2,113	24.4% (259)	27.5% (290)	$\chi^2_{(1)}$ = 2.65; *p* = .104
Q2 (item 2)	2,093	20.6% (216)	22.0% (230)	$\chi^2_{(1)}$ = .61; *p* = .435
Q3 (item 1)	2,186	31.9% (347)	31.4% (345)	$\chi^2_{(1)}$ = .04; *p* = .835
Q3 (item 5)	2,185	54.1% (589)	53.0% (581)	$\chi^2_{(1)}$ = .30; *p* = .582
Q4	2,166	5.7% (61)	6.5% (71)	$\chi^2_{(1)}$ = .61; *p* = .437
Q5	1,860	37.4% (343)	25.7% (242)	$\chi^2_{(1)}$ = 29.73; *p* < .001
Q6 (item 2)	2,136	22.6% (240)	24.5% (263)	$\chi^2_{(1)}$ = 1.06; *p* = .304
Q6 (item 4)	2,097	22.4% (234)	23.6% (249)	$\chi^2_{(1)}$ = .42; *p* = .518

Table 6.7. Non-differentiation

	Condition (A): open-ended probes	**Condition (B): without probes**	**Significance level**
Absolute non-differentiation	% (*n*)	% (*n*)	
Q3	3.3% (36)	4.1% (45)	$\chi^2_{(1)}$ = .971; *p* = .324
Q6	10.6% (116)	9.3% (103)	$\chi^2_{(1)}$ = .965; *p* = .326
Mean root of pairs	mean (std)	mean (std)	
Q3	.335 (.193)	.338 (.203)	$T_{(2198)}$ = -.435; *p* = .664
Q6	.483 (.242)	.473 (.234)	$T_{(2198)}$ = 1.046; *p* = .295

6.6. Discussion and conclusion

The purpose of the presented research was to determine whether and in which ways embedding open-ended probes into web surveys impacts the process of responding to closed survey questions. In doing so, it took a different perspective than many current studies in the area of open-ended questions, which examine contextual effects on the response quality to open-ended questions and probes.[15] The study differentiated between

[15] A series of studies have examined the impact of respondent and survey characteristics (Schmidt et al., 2020) on response quality to open-ended questions, including the size of the answer box (Meitinger &

the effects of open-ended probes on survey completion, on the survey questions the probes pertain to, and on subsequent survey questions. To this end, a randomized web survey experiment was carried out, in which closed survey questions were fielded with and without open-ended probes using a paging design. Inserting open-ended probes increased survey break-off and impacted the survey questions the probes pertained to in the form of increased backtracking and answer changes. Effects on subsequent questions occurred in single cases (see Table 6.8 for an overview of the hypotheses and results).

Table 6.8. Study 2: Summary of results

Hypotheses		**Result**
Embedding open-ended probes into web surveys…		
Impact on survey break-off		
H1	… increases survey break-off	Confirmed
Impact on preceding survey questions		
H2a	… increases backtracking	Confirmed
H2b	… increases answer changes	Confirmed
Impact on subsequent survey questions		
H3	… increases response times	Confirmed for one question only
H4	… decreases nonresponse in terms of item skipping and choosing non-substantive response options	Not confirmed
H5a	… does not impact mean scores	Confirmed
H5b	… increases extreme responding	Confirmed for one question only
H5c	… increases non-differentiation	Not confirmed

The majority of break-offs occurred in the condition with open-ended probes, particularly after the first- and second-shown probe. The open-ended probes were not announced at the beginning of the survey, which may have contributed to the increase in break-offs for the first-shown probes (rather than at the announcement that probes will be asked). Importantly, respondents with lower education and women were more likely to

Kunz, 2022; Smyth et al., 2009; Zuell et al., 2015), the use of placeholder text (Kunz, Quoß, & Gummer, 2020), the order of the closed survey questions (Hadler, 2021) and the sequence of the open-ended probes (Meitinger et al., 2018).

break off the survey. Although embedding open-ended probes did not lead to an unusually high level of survey break-off, survey researchers should consider this potential nonresponse bias when implementing open-ended probes in web surveys.

Embedding open-ended probes significantly increased backtracking and answer changes to previous survey questions. In the present study, 9% of respondents who received open-ended probes returned to a previous question, while only 3% of respondents did this in the condition without probes. Across both conditions, about one of three respondents who backtracked changed their response to the preceding survey question. Like survey break-off, backtracking and answer changes to previous questions occurred most often in response to the first open-ended probes.

Asking open-ended probes did not impact subsequent questions for the most part. There were no significant effects on mean scores, item skipping or non-differentiation for any of the examined questions. For four of the five examined questions, there were no significant effects on response time, choosing a non-substantive response option or extreme responding. This indicates that, in most cases, the cognitive processing of survey questions and subsequent web survey data are not impacted by inserting open-ended probes. However, there are notable exceptions.

Respondents were significantly more likely to choose a non-substantive response option to the single-item measure of life satisfaction (Q4) in the condition with open-ended probes. Moreover, respondents took significantly longer to answer and were more likely to give an extreme response to the question on relationship satisfaction (Q5) in the condition with open-ended probes.

There are several possible explanations for these findings. First, the responses to the open-ended probes indicated that the question on relationship satisfaction was flawed because it lacked a response category to indicate that one was currently not in a relationship. By the time respondents in condition (A) reached the survey question on relationship satisfaction, they were certainly expecting an open-ended probe to follow. Possibly, respondents lacking a suitable response option dealt with this irritation differently when they were expecting to be able to explain their response in an open-ended text field than when they were not expecting this. The majority of respondents who indicated that they were not in a relationship in the probe either chose the available non-substantive response option (“I do not want to answer this question”) or an extreme response option (i.e., “very unsatisfied”), while only one respondent in condition (A) chose to skip the question.[16]

However, alternative explanations should be considered. Perhaps, the effects of embedding open-ended probes on the response to subsequent closed survey questions are more likely to occur when there is a close connection to the preceding survey and probing questions. In the present study, the question on relationship satisfaction was directly preceded by the closely related construct on life satisfaction (Schwarz & Strack, 1991).

Probing techniques and probe design vary strongly, as do the survey questions they pertain to, and generalizing the results of any given study to all settings is not possible. In the present study, the effect of open-ended probing on the response time and behaviour of only one of the examined questions highlights that future research should establish which question, probe, and respondent characteristics determine when open-ended probes impact surrounding survey questions.

Thus, the present study has several limitations which point the way to future research. For instance, it could be that certain probe types, such as category selection probing, increase the likelihood of respondents backtracking and changing their answers more than other probes that do not prompt respondents to reconsider their survey response. Different spacing designs should be examined to understand whether asking probes after (almost) each survey question leads to other effects than spacing probes throughout the survey or only inserting one (random) probe. Future research should include other question types, such as behaviour and factual questions. Finally, future studies should employ further measures of data quality, such as test-retest reliability (Knowles et al., 1992).

In summary, embedding open-ended probes can increase survey break-off, backtracking, and answer changes to previous questions, though fortunately, none of these outcomes occurred very often in the present study. Survey researchers may omit a back button to prevent effects on previous questions in practice. Of course, effects on response behaviour to subsequent survey questions cannot be prevented technically; however, for this study such effects were seldom, and there is no reason to assume a

[16] Unfortunately, as the relationship status of respondents in condition B without probes was not determined in the course of the study, it cannot be determined whether respondents in condition B who were not in a relationship at the time of the survey systematically chose other survey response options.

worrisome impact on data collection. More than ever, Schuman’s (1966) suggestion to ask single open-ended probes to a subsample of a survey seems a timely and pragmatic compromise in order to control for rare effects of open-ended probes on response behaviour while gaining insights into respondents’ thought processes and thereby validating survey responses.

6.7. Data availability

The quantitative data set of this study and analysis file is available under the following link:

Hadler, Patricia (2023): Hadler 2023 SMR_Effect of openended probes_Analysis.sps. figshare. Dataset. https://doi.org/10.6084/m9.figshare.22499557.v1. The answers to the open-ended questions are not publicly available due to them containing information that could compromise participant privacy.

6.8. Appendix Study 2

A.1. Questionnaire

Q1. Many people use the terms "left" and "right" when referring to different political attitudes. When you think of your own political views, where would you rank those views on this scale? *[Viele Leute verwenden die Begriffe „links“ und „rechts“, wenn es darum geht, unterschiedliche politische Einstellungen zu kennzeichnen. Wenn Sie an Ihre eigenen politischen Ansichten denken, wo würden Sie diese Ansichten auf dieser Skala einstufen?]*

Response scale: 1 left, 2, 3, 4, 5, 6, 7, 8, 9, 10 right, no reply *[1 links, 2, 3, 4, 5, 6, 7, 8, 9, 10 rechts, keine Angabe]*

P1 [introduction]. We would like to receive more information on the previous question. *[Wir möchten zu der vorherigen Frage gerne noch nähere Informationen erhalten.]*

P1a. The question just asked was: [Q1]. Would you please tell me what you associate with the term "left"? *[Die Frage soeben lautete: [Q1]. Würden Sie mir bitte sagen, was Sie mit dem Begriff „links“ verbinden?]*

P1b. The question just asked was: [Q1]. Would you please tell me what you associate with the term "right"? *[Die Frage soeben lautete: [Q1]. Würden Sie mir bitte sagen, was Sie mit dem Begriff „rechts“ verbinden?]*

Q2. We are interested in how you rate the politicians in Germany. What would you say... *[Wir interessieren uns dafür, wie Sie die Politiker in Deutschland einschätzen. Was würden Sie sagen…]*

1. ... how many politicians are honest with voters? *[…wie viele Politiker sind ehrlich zu den Wählern?]*
2. ... how many politicians are in politics to achieve as much personal gain as possible? *[…wie viele Politiker sind in der Politik, um möglichst viel für sich selbst herauszuholen?]*

Response scale: Almost all, Most of them, About half, Only few, Almost none, I don't want to answer *[So gut wie alle, Die meisten, Etwa die Hälfte, Nur wenige, So gut wie keine, Möchte ich nicht beantworten]*

P2. One of the previous statements was [ITEM 1]. Your answer was [ANSWER ITEM 1]. Why did you choose this answer? / One of the previous statements was [ITEM 2]. What do you understand by "to achieve personal gain" in this question? *[Eine der vorangegangenen Aussagen lautete: [ITEM 1] Ihre Antwort lautete: [ANSWER ITEM 1]. Wieso haben Sie sich für diese Antwort entschieden? / Eine der vorangegangenen Aussagen lautete: [ITEM 2]. Was verstehen Sie in dieser Frage unter „für sich selbst etwas herauszuholen“?]*

Q3. The following statements may apply to you more or less. For each statement, please indicate how much the statement applies to you. *[Die folgenden Aussagen können auf Sie selbst mehr oder weniger zutreffen. Bitte geben Sie bei jeder Aussage an, wie sehr die Aussage auf Sie zutrifft.]*

1. It has happened that I have taken advantage of someone in the past. *[Es ist schon mal vorgekommen, dass ich jemanden ausgenutzt habe.]*

2. Even if I am feeling stressed, I am always friendly and polite to others. *[Auch wenn ich selbst gestresst bin, behandle ich andere immer freundlich und zuvorkommend.]*
3. Sometimes I only help people if I expect to get something in return. *[Manchmal helfe ich jemandem nur, wenn ich eine Gegenleistung erwarten kann.]*
4. In an argument, I always remain objective and stick to the facts. *[Im Streit bleibe ich stets sachlich und objektiv.]*
5. I have occasionally thrown litter away in the countryside or on to the road. *[Ich habe schon mal Müll einfach in die Landschaft oder auf die Straße geworfen.]*
6. When talking to someone, I always listen carefully to what the other person says. *[Wenn ich mich mit jemandem unterhalte, höre ich ihm immer aufmerksam zu.]*

Response scale: Doesn't apply at all, Doesn't apply much, Applies partially, Fully applies *[Trifft gar nicht zu, Trifft wenig zu, Trifft etwas zu, Trifft ziemlich zu, Trifft voll und ganz zu]*

P3. One of the previous statements was: [ITEM 1]. What do you understand by taking advantage of someone in this question? Please name examples. / One of the previous statements was: [ITEM 5]. Your answer was [ANSWER ITEM 5]. Why did you choose this answer? *[Eine der vorangegangenen Aussagen lautete: [ITEM 1]. Was verstehen Sie in dieser Frage darunter, jemanden auszunutzen? Bitte nennen Sie Beispiele. / Eine der vorangegangenen Aussagen lautete: [ITEM 5]. Ihre Antwort lautete: [ANSWER ITEM 5]. Warum haben Sie sich für diese Antwort entschieden?]*

Q4. The next question is about your overall satisfaction with life. How satisfied are you, all in all, with your life at present? *[Nun geht es um Ihre allgemeine Lebenszufriedenheit. Wie zufrieden sind Sie gegenwärtig, alles in allem, mit Ihrem Leben?]*

Response scale: 1 Totally unsatisfied, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11 Totally satisfied, I don't want to answer *[1 Überhaupt nicht zufrieden, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11 Völlig zufrieden, Das möchte ich nicht beantworten]*

P4. The previous question was: [Q4]. On a scale of 1 to 11, in which 1 means "Totally unsatisfied" and 11 means "Totally satisfied", you chose [ANSWER Q4]. What aspects of your life did you think about when answering the question? *[Die Frage soeben lautete: [Q4]. Auf einer Skala von 1 bis 11, wobei 1 "überhaupt nicht zufrieden" und 11 "völlig zufrieden" bedeutet, haben Sie den Wert [ANSWER Q4] angekreuzt. An welche Aspekte Ihres Lebens haben Sie beim Beantworten der Frage gedacht?]*

Q5. Denken Sie bitte einmal an Ihre partnerschaftliche Beziehung (Ehe oder Freund/in). Wie zufrieden sind Sie zurzeit mit Ihrer Partnerschaft?

Response scale: 1 Totally unsatisfied, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11 Totally satisfied, I don't want to answer *[1 Überhaupt nicht zufrieden, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11 Völlig zufrieden, Das möchte ich nicht beantworten]*

P5. The previous question was: [Q5]. On a scale of 1 to 11, in which 1 means "Totally unsatisfied" and 11 means "Totally satisfied", you chose [ANSWER Q5]. What aspects of your partnership did you think about when answering the question? *[Die Frage soeben lautete: [Q5]. Auf einer Skala von 1 bis 11, wobei 1 "überhaupt nicht zufrieden" und 11 "völlig zufrieden" bedeutet, haben Sie den Wert [ANSWER Q5] angekreuzt. An welche Aspekte Ihrer Partnerschaft haben Sie beim Beantworten der Frage gedacht?]*

Q6. The following statements deal with tasks that people may have in their family. Please indicate the extent to which you agree with each of the following statements. *[Die folgenden Aussagen beschäftigen sich mit Aufgaben, die Menschen möglicherweise in ihrer Familie haben. Geben Sie bitte jeweils an, inwiefern Sie den folgenden Aussagen zustimmen.]*

1. Parents should do everything for their children, even at the expense of their own welfare. *[Eltern sollten alles für ihre Kinder tun, selbst auf Kosten ihres Wohlergehens.]*
2. Grandparents should contribute to the economic security of their grandchildren and their families. *[Großeltern sollten zur wirtschaftlichen Absicherung ihrer Enkel und deren Familien beitragen.]*

3. Grandparents should help care for their grandchildren when they are young. *[Großeltern sollten bei der Betreuung ihrer Enkel helfen, wenn diese noch klein sind.]*
4. In order not to burden their children, parents in need of care should seek care in a home. *[Um ihre Kinder nicht zu belasten, sollten sich pflegebedürftige Eltern in einem Heim betreuen lassen.]*
5. Adult children should take their parents in their old age. *[Erwachsene Kinder sollten ihre Eltern im hohen Alter bei sich aufnehmen.]*
6. Adult grandchildren should assist in the care and nurturing of their grandparents. *[Erwachsene Enkelkinder sollten bei der Betreuung und Pflege ihrer Großeltern mithelfen.]*

Response scale: Completely agree, Partially agree, Neither agree nor disagree, Partially disagree, Complete disagree, I don't want to answer *[Stimme voll zu, Stimme etwas zu, Weder/noch, Lehne etwas ab, Lehne ganz ab, Möchte ich nicht beantworten]*

P6. One of the previous statements was: [ITEM 4]. You answer was: [ANSWER ITEM 4]. Why did you choose this answer? / One of the previous statements was: [ITEM 2]. What do you understand by „economic security“ in this question? Please name examples. *[Eine der vorangegangenen Aussagen lautete: [ITEM 4]. Ihre Antwort lautete: [ANSWER ITEM 4]. Wieso haben Sie sich für diese Antwort entschieden? / Eine der vorangegangenen Aussagen lautete: [ITEM 2]. Was verstehen Sie in dieser Frage unter „wirtschaftlicher Absicherung“? Bitte nennen Sie Beispiele.]*

A.2. Sample composition

	Condition (A): with open-ended probes	**Condition (B): without probes**	**Significance level**
Gender	% (*n*)	% (*n*)	
Male	51.3% (562)	50.6% (559)	
Female/non-binary	48.7% (534)	49.4% (545)	$\chi^2_{(1)}$ = 0.091; *p* = .763
Education			
Low	27.8% (305)	28.6% (316)	
Medium	31.0% (340)	31.0% (342)	
High	41.1% (451)	40.4% (446)	$\chi^2_{(2)}$ = 0.199; *p* = .905
Region in Germany			
Former West Germany	78.9% (865)	79.8% (881)	
Former East Germany	21.1% (231)	20.2% (223)	$\chi^2_{(1)}$ = 0.259; *p* = .611
Device used			
PC/Laptop	73.3% (803)	72.4% (799)	
Smartphone	26.7% (293)	27.6% (305)	$\chi^2_{(1)}$ = 0.222; *p* = .638
Age	mean (std)	mean (std)	
Mean age (in years)	44.5 (14.61)	45.63 (15.14)	$T_{(2198)}$ = -1.780; *p* = .075

7. STUDY 3: THE IMPACT OF QUESTION ORDER ON PROBES

A version of this chapter has been published as:
Hadler, Patricia (2021). Question order effects in cross-cultural web probing: Pretesting behavior and attitude questions. *Social Science Computer Review, 39*(6), 1292-1312. DOI: 10.1177/0894439321992779

7.1. Introduction

Pretesting questionnaires before fielding is considered indispensable by both elementary textbooks and experienced researchers in order to ensure data quality (Presser et al., 2004; Presser & Blair, 1994). Cognitive pretesting asks respondents to verbalize their thought processes while answering survey questions (Beatty & Willis, 2007). The method is used to evaluate respondents' answer process, identify problems they encounter while answering survey questions, and, based on these findings, suggest question revisions prior to data collection (Willis, 2015a).

Until recently, cognitive pretesting took place mainly in the form of face-to-face interviews (Willis, 2005). In a cognitive interview, a participant is presented a survey question by an interviewer and answers it. This is followed by one or several "specific questions or probes about how the participant set about answering the question being tested" (Collins, 2015, p. 14). For instance, a respondent may answer a survey question on their past behaviour. Subsequent probing questions might ask how the respondent remembered their past behaviour or what type of activities they included in their answer. Once the probes have been answered, the interviewer and participant move on to the next survey question and probes relating to this question. If the next question is an attitude question on the same topic, probing questions may be used to understand why the respondent agreed or disagreed with the statement or how they came to their answer.

In the past decade, web probing has developed as a complementary self-administered form of cognitive pretesting (Behr et al., 2012b; Behr, Braun et al., 2012). Implementing probing techniques from cognitive interviewing into web surveys has proven especially useful to test web surveys in the same mode (Fowler & Willis, 2020, p. 466) and has gained popularity in the context of cross-national pretests (Behr et al., 2020; Braun, Behr, Kaczmirek, & Bandilla, 2014). Online probes are open-ended questions that directly relate to a foregoing (usually closed) survey question (Behr et al., 2017). Web probing encounters higher levels of nonresponse and generates shorter probe responses than cognitive interviews (Lenzner & Neuert, 2017; Meitinger & Behr, 2016). However, previous research indicates that both modes generate similar findings (Meitinger & Behr, 2016) and lead to similar revisions (Lenzner & Neuert, 2017).

Regardless of mode, the focus of cognitive pretesting lies on testing individual questions rather than the questionnaire as a whole (Lenzner et al., 2016). Cognitive pretests take place at an early stage in questionnaire development and usually examine only parts of questionnaires. In consequence, question sequence in cognitive pretesting must not necessarily be identical to that in the later survey. Ultimately, analysis of cognitive pretesting rests on the assumption that pretest results are independent of a tested question's position during pretesting.

Although cognitive pretesting has been used in the past to understand the causes of question order effects (Bishop et al., 1985; Bishop, 1992), the reverse effect of question order on cognitive pretesting has yet to be examined. To date, we do not know whether the order of presenting survey questions affects the responses given to probing questions and, if this is the case, whether this can impact revisions to survey questions. This research gap is all the more surprising, as the order of presenting questions can influence the response to survey questions (e.g., Schwarz & Sudman, 1992), and the response to survey questions is in turn known to influence respondents' likelihood of giving substantive probe responses (Zuell & Scholz, 2015). Should probe responses be impacted by the sequence of survey questions in a cognitive pretest, this endangers the validity of cognitive pretest results and, in consequence, the quality of the final questionnaire and survey data.

Question order effects "occur most when two questions are asked sequentially on the same topic or very similar topics" (Stark et al., 2020, p. 28; Tourangeau et al., 2003). One such case is the presentation of behaviour and corresponding attitude questions. Questionnaires frequently include both question types, as many psychological frameworks use attitudinal measures to explain or predict behaviour (most prominently in the theory of reasoned action; Ajzen & Fishbein, 1980; and the theory of planned behaviour; Ajzen, 1985). Attitude-behaviour consistency is generally moderate (Ajzen & Fishbein, 1977) and can be impacted by question order (Budd, 1987).

Even well-documented question order effects vary across countries. Reasons for differences in question order effects between countries are manifold, including substantive, topic-specific differences between countries but also differences in survey response styles (Stark et al., 2020). Moreover, past research on web probing has repeatedly pointed to cross-cultural differences in probe response quality irrespective of survey response behaviour (e.g., Meitinger et al., 2019).

The research presents an experimental study on the effects of question order of a behaviour and attitude survey question on probe responses in cross-cultural web probing. The following section discusses the potential influence of question order on probe response content and consistency. The topic of fare evasion is introduced as an example of related behaviour and attitude questions. Next, the research design is presented, followed by the results. The discussion outlines the implications for web probing methodology.

7.2. Question order and probe responses

Examining the impact of question order on open-ended probes differs in several ways from analysing closed survey questions. For one, responses to open-ended questions are not limited by predefined answer categories (Reja et al., 2003). Moreover, examining order effects for probes is more complex as it involves not just two questions but at least four: two survey questions and two probes. On the other side, surveys and cognitive pretests have in common that they are founded on the dialogue structure of asking and answering and the reliance on contextual information to interpret questions and give meaningful answers (Conrad et al., 2014).

Probes can directly follow the survey question they pertain to (concurrent or embedded probing) or be administered after several or all survey questions (retrospective probing; Collins, 2015, p. 120). The rationale behind concurrent probing is that respondents are better able to recall their thoughts while answering the survey question and do not post-rationalize their survey responses. Retrospective probing can be used to prevent probes from influencing thought processes during and responses to subsequent survey questions. To date, there is little research comparing the effects of concurrent and retrospective probing on probe responses (for a notable exception, see Fowler & Willis, 2020), nor in how far concurrent probing of related survey questions impacts responses to subsequent survey and probing questions. This study focuses on concurrent probing, where such question order effects are considered more likely to occur and endanger probe response quality.

The following sections apply two main perspectives of examining order effects to the context of cognitive pretesting. The first perspective focuses on the single probe and the occurrence of probe response content as a function of question order. The second analytical perspective examines the relation of responses to one another in the form of reported attitude-behaviour consistency. The third section adds a cross-cultural perspective to question order effects in general and particularly web probing.

7.2.1. Probe response content

In its simplest form, testing for question order effects can be carried out by comparing responses as a function of question order. This main effect of question order on the response to the later shown question has been coined unconditional order effect (Rasinski et al., 2012). In contrast, conditional order effects do not automatically occur when a question is preceded by another question, but only when the respondent gives a certain answer to that preceding question. Thus, conditional order effects take into account a possible "interaction between question order and response to the antecedent question" (Smith, 1992, p. 164).

Imagine that a survey question on a behaviour is asked, followed by a probe. This, in turn, is followed by a survey question on the attitude toward that behaviour and another probe. The Gricean (1975) maxim of relation expects respondents to offer relevant information to these probes. For instance, a respondent is asked a behaviour question on fare evasion, such as "Have you ever used public transport without having a valid ticket?" The respondent is asked to explain her answer in an open-ended probe. The answer to this probe should pertain to their behaviour (i.e., "I have never avoided the fare and always buy a ticket"). A subsequent attitude question may ask how she feels toward people avoiding the fare. The response to a probe following this question should relate to her attitude (i.e., "avoiding the fare is absolutely wrong because other people pay the price").

The order of presenting the two survey and subsequent probing questions may impact probe responses. When respondents are first presented the behaviour survey question and probe, they may elaborate on both their behaviour and attitude. When respondents are subsequently presented the attitude survey question and probe, they might be less inclined to elaborate on their attitude a second time. Such an effect could be caused by another conversational maxim, known as the norm of non-redundancy (Clark & Haviland, 1977), which requires information provided by each party to be new, not reiterating information the recipient already has (Schwarz, 1995). This situation describes a conditional order effect, in which the occurrence of probe response content is a function of both question sequence and whether this content has already been mentioned in answer to the first-shown probe. More precisely, in the abovementioned scenario, attitude-related probe response content should decrease when the behaviour survey question is presented first, and the respondent already mentioned their attitude in response to the anteceding probe.

However, researchers have also considered the notion of unconditional question order effects on probe responses, with pretest participants being inclined to "provide less information to later probes because they *believe* they have already provided relevant information" (Fowler & Willis, 2020, p. 463, italics by author). In this scenario, respondents would be less likely to mention content in answer to the second-shown probe regardless of which content they included in response to the first-shown probe. Such an effect would be particularly problematic if content indicating problems with a survey question are not uncovered, such as difficulties understanding a term used in the question, retrieving the information required, or finding a suitable answer category.

The applicability of the communicative principles underlying potential question order effects has not been tested in a web probing context. However, web probing implements the same or very similar probing techniques and sequence as cognitive interviewing, implicitly assuming similar communicative principles. Following this notion, the possibility of unconditional and conditional question order effects leads to two hypotheses on the occurrence of probe response content in answer to the second-shown probe in web probing:

H1: When first one survey question and subsequent probe and then a second survey question and subsequent probe on a similar topic are presented, probe response content is less likely to be mentioned in answer to the second-shown probe (unconditional question order effect).

H2: When first one survey question and subsequent probe and then a second survey question and subsequent probe on a similar topic are presented, probe response content is less likely to be mentioned in answer to the second-shown probe *if this content has already been mentioned in answer to a previous probe* (conditional question order effect).

7.2.2. Consistency of probe responses

Another important aspect when examining question order effects is how responses relate to one another. In studies examining general and specific questions, Schwarz, Strack, and Mai (1991) found that the correlation between marital and life satisfaction differed depending on question order when the two questions were assigned to the same conversational context. Studies examining the relation of two or more questions to one another have coined the term associational question order effect (Rasinski et al., 2012).

The relationship of attitude and behaviour measures is often described in terms of consistency. Consistent responses are when attitudes and behaviour correspond to each other, for instance when respondents hold a positive attitude toward behaviour that they (intend to) engage in and negative attitudes toward behaviour they do not (intend to) engage in. Reprising the example above, holding a lenient attitude toward fare evasion would be consistent for respondents who commit fare evasion.

Many factors influence the strength of relationship between behaviour and attitude questions, such as strength of the attitude, visibility of the behaviour, and psychological state of the respondent (Fazio, Zanna, & Cooper, 1978; Liska, 1984). Also, respondents with past personal experience with a behaviour are more likely to give consistent self-reports (Fazio & Zanna, 1981; Zanna, Olson, & Fazio, 1981).

Regarding question order, consistency between behaviour and attitude self-reports increases when behaviour-related questions are asked first (Budd, 1987). Respondents are likely to align their attitudinal responses with self-reports on behaviour, especially when they establish a normative principle between two questions (Smith, 1992). An explanation for this effect is that attitudinal measurement instruments can prompt in situ attitude formation rather than assessing pre-existing attitudes. Preceding questions are a possible source of accessible information for context-dependent questions (Bless & Schwarz, 2010). Respondents may base attitude judgments on information about their reported behaviour when the survey context supports this (Schwarz & Bohner, 2001). Based on the notion of cognitive consistency (Heider, 1958), when attitudes are created or changed, this is mostly done in a way that is consistent with behaviour, thus reducing cognitive dissonance (Festinger, 1957) and enhancing self-perception (Bem, 1967).

This leads to the following hypothesis concerning probe responses:

H3: Consistency of probe responses is higher when the behaviour question is asked first.

7.2.3. Cross-cultural question order effects in web probing

Much research on question order effects has examined one country only. However, even classic question order effects vary across countries (Stark et al., 2020). For instance, question order effects on the strength of relation between general and specific questions have been found across a range of topics such as marital and general satisfaction (Schwarz, Strack, & Mai, 1991) and customer satisfaction (Schul & Schiff, 1993). A study on academic and general satisfaction reproduced the effect in a German sample but could not replicate it with Chinese participants (Haberstroh, Oyserman, Schwarz, Kühnen, & Ji, 2002). In the aftermath, the applicability of the underlying conversational norms (Grice, 1975) to collectivist cultures has been questioned (Schwarz, Oyserman, & Peytcheva, 2010). In a comparative study involving 11 countries, Stark et al. (2020) find that both differences in survey response styles and topic-specific differences contribute to explaining differing question order effects across countries.

Topic-specific differences between countries may impact question order effects for attitude and behaviour questions if, for instance, a certain behaviour is more common or considered to be more acceptable in one country. Returning to the example of fare evasion, Germany and the United States have very different usage levels of public transport and dominant control systems for fare evasion (Buehler, 2011; Buehler & Pucher, 2012). In the United States, most public transport stations have paid areas that are physically secured in the form of ticket gates. In contrast, Germany relies on an honour-based proof-of-payment system in which passengers can board public transport without prior ticket control. Random ticket inspections and fines are used to enforce payment (Fürst & Herold, 2018). Research on public transport shows that revenue loss as a consequence of fare evasion is a financial risk in proof-of-payment systems, as boarding without a valid ticket is much easier, resulting in higher levels of fare evasion (Barabino, Salis, & Useli, 2014). In the present case, cross-cultural differences in self-reported behaviour of fare evasion may result in differences in the attitude-behaviour relationship and impact question order effects.

Regardless of cross-cultural differences in survey responses, web probing experiments have repeatedly revealed differences in probe response quality even between post-industrial, individualist countries. Meitinger et al. (2019) found that American respondents generate higher levels of probe nonresponse, provide fewer themes, write shorter responses, and take less time to respond than German respondents. In another study, Meitinger et al. (2018) found that the sequence of asking several probes following one survey question impacted response quality, with respondents from different countries varying whether and in which way their response quality decreased. While U.S. respondents were likely to react with probe nonresponse, German respondents were hardly affected by probe sequence. Thus, U.S. respondents generally show lower probe response quality than German respondents, and probe response quality seems to be impacted more strongly in the United States by contextual factors. However, these studies operationalized probe response quality using measures such as probe nonresponse or length of response, and it remains unclear how the prevalence of probe response content or consistency might differ between countries as a function of question order.

In summary, the effects of question order on probe response content and consistency are likely to differ across countries, both due to differences in response to the survey questions and due to differing probe response quality between countries. The present study defines cross-cultural differences as those differences between countries which cannot be explained by differences in survey response behaviour. An undirected hypothesis is formulated to account for differences between countries:

H4: Content and consistency of probe responses differ by country.

7.3. Procedure

7.3.1. Research design

An experimental setup was chosen which randomized the order of a behaviour and an attitude survey question and the subsequent probing questions in an online questionnaire.

A general probe was used following each survey question. General probes ask respondents to explain their thoughts while answering; thus, the wording could be kept consistent across both survey questions. In the first experimental condition, respondents were first shown the behaviour survey question (Q_beh), directly followed by the probe (P_beh). They were then presented the attitude survey question (Q_att) and subsequent probe (P_att). In the second condition, the order of the questions was reversed, with respondents first answering the attitude survey question (Q_att) and probe (P_att), and thereafter the behaviour survey question (Q_beh) and probe (P_beh). Figure 7.1 illustrates the two experimental conditions.

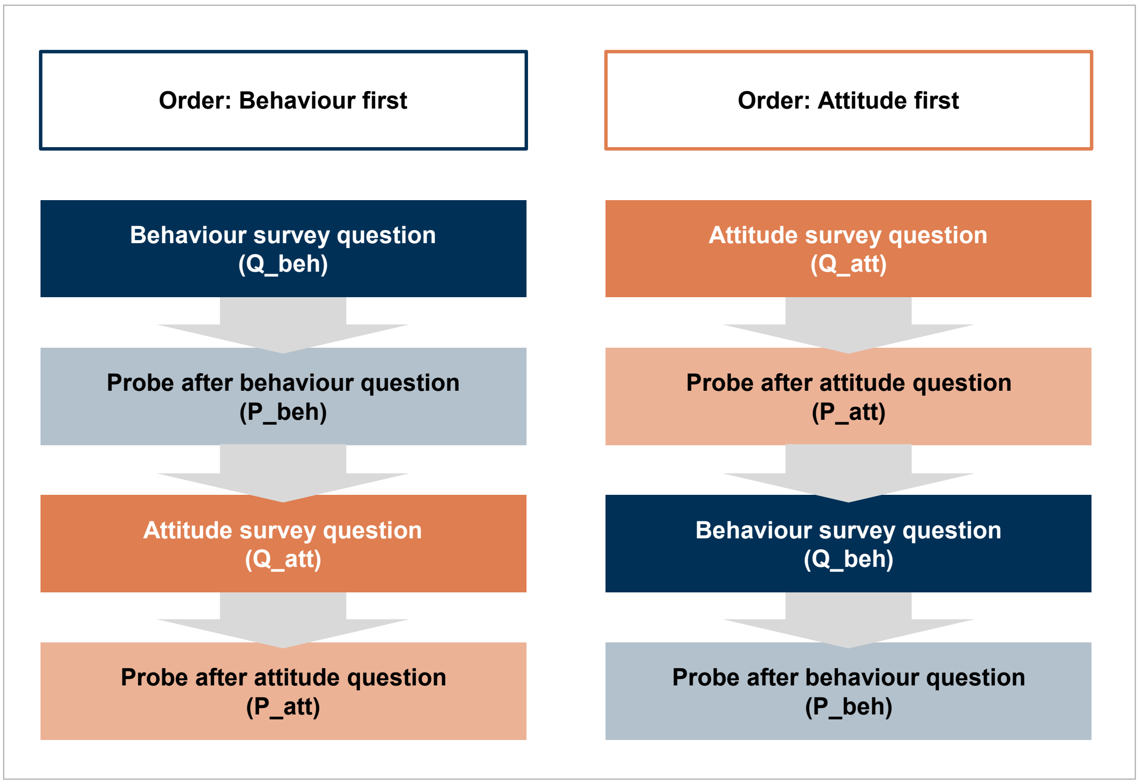

Order: Behaviour first
Order: Attitude first
Behaviour survey question (Q_beh)
Probe after behaviour question (P_beh)
Attitude survey question (Q_att)
Probe after attitude question (P_att)
Attitude survey question (Q_att)
Probe after attitude question (P_att)
Behaviour survey question (Q_beh)
Probe after behaviour question (P_beh)

Figure 7.1. Experimental design

The probing question was presented on a separate screen that repeated the question text as well as the respondent’s survey response and asked them to state what they had thought about while answering the question in an open text field. The topic examined was fare evasion. The behaviour question was drawn from the German General Social Survey (ALLBUS, 2000) and comprised four delinquent behaviours. Respondents were asked how often they had avoided the fare in the past. They could choose between six frequencies ranging from “never” to “more than 20 times” and a “don’t remember” option. A general evaluative question on a range of behaviours from World Values Survey Wave 6 (2014) was chosen for the attitude question. Item sequence was adjusted to show fare evasion as the first topic. Respondents were asked to indicate on a 10-point scale in how far they felt that fare evasion could be justified. (Table A.1 in the Appendix shows the wording of the survey and probing questions.)

A web probing study with respondents from Germany and the United States was conducted between July 25th and August 7th, 2018. The main panel provider was Respondi AG, based in Germany, who cooperated with an international partner to recruit the U.S. sample. Equal quotas for gender, age (18-29, 30-49, and 50-64 years), and education (lower and higher) were used. Of the 1,947 panelists who responded to the survey invitation, 1,248 were screened out and 400 (Germany: *n* = 192; United States: *n* = 208) completed the survey, resulting in a completion rate of 57% (Callegaro & DiSogra, 2008). The web probing study contained a range of experiments. The questions analysed in this study were from the middle of the survey and were shown directly after each other. Respondents who did not answer any of the open-ended probing questions in the course of the web survey (i.e., neither in the reported experiment nor in any part of the survey) were excluded from the sample, resulting in a final sample of *n* = 333 respondents (Germany: *n* = 167; United States: *n* = 166). A chi-square test of independence revealed no significant association between the experimental condition and gender, age, or education in either country (Table A.2 in the Appendix). The average survey completion time of the total survey was 27.6 min in Germany and 27.4 min in the United States for the final sample. Respondents were paid 2.00 Euros as an incentive for survey completion.

7.3.2. Coding scheme

In order to quantify probe response content and consistency, probe responses were coded using three coding schemes. These codes served as dependent variables in the analyses. The first two coding schemes relate to probe response content (Hypotheses 1 and 2), the third to consistency (Hypothesis 3). The first scheme indicated whether and what type of behaviour- and attitude-related content was contained in the probes. The second scheme was problem-based and indicated whether respondents reported a questionnaire understanding problem with the behaviour survey question. The third scheme indicated whether responses to the two probes were consistent to one another.

Coding Scheme 1 was created by the author. After the initial coding scheme had been developed, the author and a research assistant coded one of the probes. Following an evaluation of differences, the scheme was refined to comprise more elaborate categorization rules. Then, both coders coded all probe responses. Both probes were coded once for behaviour- and attitude-related content. Cohen's k was strong, with values between .823 and .916 (Table A.3 in the Online Appendix). Differences in coding were discussed and the final codes assigned mutually. Table 7.1 gives an overview of all codes.

Coding Scheme 1: Behaviour- and attitude-related content. For behaviour-related probe response content, the code indicated whether a respondent explicitly admitted to having avoided the fare in the past (1 = admit) or not (2 = do not admit), either by explicitly negating or remaining vague about their past behaviour. If the probe response made no reference to the respondent's past or present personal behaviour, it was coded as 0 = no mention of behaviour. For attitude-related probe response content, the code distinguished between absolute condemnation (1 = absolute condemnation) and a more lenient attitude (2 = lenient attitude). If the probe response did not contain any information on the respondent's personal attitude toward fare evasion, it was coded as 0 = no mention of attitude. Responses that contained neither behaviour- nor attitude-related content were coded as non-substantive. Answers in this category included nonresponse, single characters, off-topic remarks, and other non-codable content. The codes from Coding Scheme 1 were further recoded into binary variables to mark the presence of behaviour- and attitude-related probe response content. One code indicated whether a response mentioned the respondent's past or present personal behaviour (1 = yes, 0 = no) and a second code whether it mentioned the respondent's attitude in any form (1 = yes, 0 = no).

Table 7.1. Coding schemes

Coding Scheme	Value	Description and examples
Coding Scheme 1: Behaviour- and attitude-related content	**Behaviour**	
	1 Explicitly admit	Explicit mention of avoiding the fare at least once in the past (deliberate or unintentional, also avoiding part of the fare): – “The times that I got on without paying” – “Once, when I had no money”
	2 Does not explicitly admit	Clear statement that respondent has never avoided the fare, neither deliberately nor by accident – “I never did this” – “I rarely use public transport, but when I did I always paid” Response refers to behaviour, but remains unclear as to the action – “I thought about what I did”
	0 No mention of behaviour	Probe response does not mention the respondent’s past or present personal behaviour
	Attitude	
	1 Absolute condemnation	Respondent does not approve of behaviour, indicating that there is no justification – “If you can't pay, don't use public transport” – “Avoiding the fare is stealing”
	2 More lenient attitude	Respondent finds the behaviour justifiable / explainable, at least under certain conditions – “It depends on…” – “Sometimes you forget or don’t have enough money”
	0 No mention of attitude	Probe response does not give any evaluative judgment on fare evasion
Coding Scheme 2: Problem-related content	1 named	Respondent clearly states that they have never used public transport before, and are missing an answer option to indicate this "I’ve never had to take public transportation”
	0 not named	Respondent does not (clearly) indicate that they have never used public transport before

(Table 7.1 continued)

Coding Scheme	Value	Description and examples
Coding Scheme 3: Consistency of probe responses	1 Consistent	Respondent admits to having avoided the fare in answer to P_beh (theme code 1, behaviour = 1) and displays a more lenient attitude towards the topic of fare evasion in answer to P_att (theme code 1, attitude = 2) OR Respondent does not explicitly admit to having avoided the fare in the past (theme code 1, behaviour = 2) in answer to P_beh and displays absolute condemnation towards this behaviour in P_att (theme code 1, attitude = 1)
	2 Inconsistent	Respondent admits to having avoided the fare in answer to P_beh (theme code 1, behaviour = 1) and displays absolute condemnation towards this behaviour in P_att (theme code 1, attitude = 1) OR Respondent does not explicitly admit to having avoided the fare in the past in P_beh (theme code 1, behaviour = 2) and displays a more lenient attitude towards the topic of fare evasion in P_att (theme code 1, attitude = 2)
	3 Unascertainable	Respondent does not relate to his/her behaviour in response to P_beh OR does not relate to his/her attitude in response to P_att

Note. Coding schemes 1 and 3 were carried out for each probe separately; Coding scheme 2 was applied to the probe following the behaviour question (P_beh).

Coding Scheme 2: Problem-related content. The second coding perspective focused on problems respondents reported while answering the survey questions and pointed to issues of question design. Problems were classified along the cognitive process of survey response (Tourangeau et al., 2000) for each survey question separately. Problems included misunderstanding of the term public transport (i.e., to include airplanes or taxis) and the concept of fare evasion (i.e., whether or not to include unintentional fare evasion).[17] The present analysis includes the only problem that emerged with sufficient size for quantitative analysis. This issue pertained to the behaviour survey question (Q_beh) and pointed to a missing answer category to indicate that a respondent had never used public transport before (1 = 'answer category missing'

[17] The coding scheme is available from the author on request.

mentioned, 0 = not mentioned). This response option is crucial to distinguish between honest customers and people to whom the question does not apply.

Coding Scheme 3: Consistency of probe responses. The third coding scheme coded probe responses as consistent (1 = consistent) if the respondent (a) admitted to having avoided the fare in the past and displayed a lenient attitude toward this behaviour or (b) did not admit to the behaviour and reported absolute condemnation. Probe responses were coded as inconsistent (2 = inconsistent) if the respondent (c) admitted to the behaviour but displayed absolute condemnation as an attitude or (d) did not admit to the behaviour but displayed a lenient attitude. If probe responses did not contain both behaviour- and attitude-related content, consistency was coded as unascertainable (3 = unascertainable).

7.3.3. Data analysis

Binary logistic regression models were carried out to examine the occurrence of probe response content and a multinomial logistic regression to examine the consistency of responses to one another. The dependent variables regarding the occurrence of probe response content were the binary behaviour and attitude variables from Coding Scheme 1 and the problem-related content from Coding Scheme 2. The dependent variable for the consistency of probe responses was Coding Scheme 3.

Question order (1 = behaviour question first, 0 = attitude question first) was used as a main predictor to test Hypotheses 1 and 3. To test whether previously mentioned probe response content is less likely to be mentioned in response to the second-shown probe (Hypothesis 2), two dummy variables were created that signified that the respective content had already been mentioned in answer to the previously shown probe (1 = probe response content mentioned previously, 0 = probe response content not mentioned previously). One dummy variable indicated this for behaviour-related, the other for attitude-related content. As problem-related content was only coded for the behaviour survey question, no dummy variable was created for the previous occurrence of problem-related probe response content. Country was inserted as a main predictor (1 = Germany, 0 = United States) to examine cross-cultural differences (Hypothesis 4).

As past research has demonstrated the impact of survey response on responses to open-ended questions (Zuell & Scholz, 2015) and question order effects may be impacted by survey response behaviour (Stark et al., 2018), models controlled for the response to the preceding survey question. To this end, survey responses were recoded into binary variables using the same logic as in Coding Scheme 1 (behaviour question Q_beh: 1 = admit, 0 = do not admit; attitude question Q_att: 1 = lenient attitude, 0 = absolute condemnation). Gender (1 = women, 0 = men) and age were included as covariates as fare evasion has been associated with young men (Cools, Fabbro, & Bellemans, 2018). Past studies have linked the strength of question order effects to lower education (Narayan & Krosnick, 1996), though this has been disputed in more recent studies (Stark et al., 2020; education: 1 = low, 0 = high). There is no previous research indicating whether the tendency toward social desirability responding impacts the likelihood of responding to probing questions. As the topic of fare evasion is potentially sensitive, the short scale for social desirability responding (KSE-G; Kemper et al., 2014) with the two dimensions "exaggerating positive traits" and "minimizing negative traits" were included as metric covariates in all models.

Prior to analysis, the prerequisites of logistic regression were tested. The multicollinearity diagnostic for metric and dichotomous predictors revealed good results, with all variance inflation factors (VIFs) slightly above 1. All models had few outliers. Data analysis was carried out using SPSS Version 24.

7.4. Results

The responses to the two survey questions confirm the reported difference in the prevalence of the fare evasion between Germany and the United States. While 52% of German respondents (*n* = 87) admitted to fare evasion, this was only the case for 15% of American respondents (*n* = 25). Interestingly, 61% (*n* = 53) of German respondents displayed absolute condemnation of fare evasion when they were first asked about their attitude; this value decreased to 40% (*n* = 32) when they were first asked about their personal fare evasion behaviour. American respondents generally showed a more lenient attitude toward fare evasion (63%; *n* = 105).

7.4.1. Probe response content

The first focus of the analysis was the occurrence of probe response content. To examine this, probe responses were coded to indicate behaviour-, attitude- and problem-related content. Table 7.2 shows the occurrence of probe response content and responses to the survey questions.

A clear majority of respondents mentioned their personal behaviour in answer to the probe following the behaviour question (P_beh: 62%; *n* = 206) and their attitude in response to the probe following the attitude question (P_att: 72%; *n* = 240). In no cases did respondents' probing answers contradict their survey responses (i.e., no respondent claimed to have committed fare evasion in answer to the survey question, but not the probing question or the other way around). Problem-related content was mentioned in answer to the probe following the behaviour question and only by U.S. respondents (P_beh: 12%; *n* = 39).

Hypothesis 1 predicted that content was less likely to be mentioned in response to a probe when the respective probe was shown second. Hypothesis 2 specified that this would be the case when the respective content had already been mentioned in response to the first-shown probe. Hypothesis 4 predicted differences between countries regarding the occurrence of content. These hypotheses were tested using binary logistic regression models for behaviour-related content to P_beh (Model 1), attitude-related content to P_att (Model 2), and problem-related content to P_beh (Model 3). To test Hypothesis 1, question order served as a main predictor. To test Hypothesis 2, the dummy variables on previous mention of probe response content were the main predictors for Models 1 and 2. To test Hypothesis 4, country was included as a main predictor in Models 1 and 2. Covariates were included as described under Data Analysis. The results of the binary logistic regressions are shown in Table 7.3.

Model 1 showed significant effects of response to the behaviour question and gender on the occurrence of behaviour-related probe response content in answer to the probe following the behaviour question. Respondents who had committed fare evasion in the past were more likely to offer behaviour-related probe response content, as were women. In contrast, respondents who reported that they had never avoided the fare and men were more likely to insert non-substantive responses (such as "fare evasion" or "nothing") or not respond to the probe at all. Contrary to Hypotheses 1 and 2, neither question order nor previous mention of behaviour-related content influenced the occurrence of behaviour-related content. However, the data basis for testing Hypothesis 2 was small, as only few respondents mentioned their personal behaviour in answer to the probe following the preceding attitude question (see Table 7.2).

Table 7.2. Response distributions for survey and probing questions

	Germany		U.S.	
	Q_beh first	**Q_att first**	**Q_beh first**	**Q_att first**
	% (n)	% (n)	% (n)	% (n)
Behaviour survey question (Q_beh)				
Admit to fare evasion	54% (43)	51% (44)	16% (11)	15% (14)
Do not admit	46% (37)	49% (43)	84% (59)	85% (82)
Attitude survey question (Q_att)				
Absolute condemnation	40% (32)	61% (53)	37% (26)	36% (35)
More lenient attitude	60% (48)	39% (34)	63% (44)	64% (61)
Probe following behaviour question (P_beh)				
Behaviour-related content (Coding Scheme 1)				
Admit	45% (36)	38% (33)	10% (7)	8% (8)
Do not admit	13% (10)	26% (23)	57% (40)	51% (49)
No behaviour-related content	43% (34)	36% (31)	33% (23)	41% (39)
Attitude-related content (Coding Scheme 1)				
Absolute condemnation	5% (4)	9% (8)	3% (2)	5% (5)
More lenient attitude	34% (27)	28% (24)	7% (5)	3% (3)
No attitude-related content	61% (49)	63% (55)	90% (63)	92% (88)
Problem-related content (Coding Scheme 2)				
“Answer category missing” mentioned	0% (0)	0% (0)	33% (23)	17% (16)
Not mentioned	100% (80)	100% (87)	67% (47)	83% (80)
Probe following attitude question (P_att)				
Attitude-related content (Coding Scheme 1)				
Absolute condemnation	26% (21)	48% (42)	30% (21)	30% (29)
More lenient attitude	45% (36)	26% (23)	43% (30)	40% (38)
No attitude-related content	29% (23)	25% (22)	27% (19)	30% (29)
Behaviour-related content (Coding Scheme 1)				
Admit	4% (3)	6% (5)	3% (2)	1% (1)
Do not admit	3% (2)	3% (3)	7% (5)	3% (3)
No behaviour-related content	94% (75)	91% (79)	90% (63)	96% (92)

Table 7.3. Probe response content, binary logistic regressions

	(1) Behaviour-related content	**(2) Attitude-related content**	**(3) Problem-related content**
N	333	333	166
	OR	OR	OR
Question order (1=behaviour first)	0.98	0.71	2.27*
Previous mention of content			
Behaviour mentioned previously (1=yes)	2.05	-	-
Attitude mentioned previously (1=yes)	-	4.03*	-
Country (1=Germany)	0.81	1.05	-
Survey response			
Behaviour question (Q_beh; 1=admit)	2.24**	-	-
Attitude question (Q_att; 1=lenient attitude)	-	1.14	-
Gender (1=women)	1.70*	2.02**	3.07*
Age	1.05	1.08	1.16*
Education (1=low)	0.74	0.68	1.04
Exaggeration of positive qualities	1.01	0.96	0.90
Under-exaggeration of negative qualities	0.75	0.96	0.60
Constant	1.52	1.69	0.13
Model χ^2 (df)	$\chi^2(9)=21.05$; *p* = .012	$\chi^2(9)=21.36$; *p* = .011	$\chi^2(6)=21.53$; *p* = .001
Nagelkerke R^2	.08	.09	.18

OR = odds ratio; * *p* < .05; ** *p* < .01; *** *p* < .001.

Model 2 showed a significant effect of having previously mentioned attitude-related probe response content (B = 1.39, standard error [SE] = .58, odds ratio [OR] = 4.03, *p* < .05) on the likelihood of doing so in answer to the probe following the attitude question. However, the direction of the effect was contrary to Hypothesis 2, with respondents who had previously volunteered attitude-related content being more likely to do this a second time. For instance, one respondent wrote in answer to the probe following the behaviour question "It's not ok to not pay" and gave a very similar response in answer to the probe following the attitude question ("not paying for the bus is like stealing"). Further, there was an unexpected difference between countries regarding the independent variable. While almost 40% of German respondents offered attitude-related content in the probe response following the behaviour question (P_beh), only 10% of U.S. respondents did this (see Table 7.2). In other words, the data basis for examining Hypothesis 2 was good for German respondents but again rather weak in the case of U.S. respondents. Gender of the respondent showed a significant effect in the model, with women more likely to mention their attitude than men. Question order had no significant effect. Thus, neither hypotheses 1 nor 2 are supported for attitude-related content.

Model 3 examined the likelihood of respondents indicating a problem responding to the behaviour survey question, namely, that a suitable response category was missing to indicate that they had never used public transport. When the behaviour question was shown first, 33% (*n* = 23) of U.S. respondents mentioned this issue in answer to the probe; when the attitude question was shown first, the rate sunk to 17% (*n* = 16; see Table 7.2). Several independent variables had to be omitted from this model. First, the problem-related content referred to the behaviour survey question and was only coded for the probe following this question (P_beh). Thus, the dummy variable regarding the previous mention of the same content was not coded, and Hypothesis 2 could not be tested in this model. Second, the problem code was only detected in the U.S. sample, certainly due to the stronger use of public transport in Germany. Thus, while the data demonstrate differences between countries, the analysis was only carried out using the U.S. sample and the variable country was omitted from the list of predictors. Finally, the problem was only coded for respondents who had answered the behaviour survey question (Q_beh) with “never.” The variable response to the survey question (Q_beh) was excluded from the model as it showed no variance.

Despite the low case number regarding the dependent variable and sample, Model 3 showed a good fit. It also showed significant effects of question order, gender, and age on the likelihood of reporting the problem. Respondents were more likely to mention this problem with the question when the behaviour question was asked first (B = 0.82, SE = .40, OR = 2.27, $p < .05$), supporting Hypothesis 1. Women and older respondents were more likely to report the problem.

In summary, question order impacted the occurrence of probe response content for Model 3 only, lending limited support for Hypothesis 1. Hypothesis 2 could not be confirmed. Indeed, for attitude-related probe response content (Model 2), the likelihood of mentioning this content even increased among respondents who had already done so previously. Hypothesis 4 was partially confirmed. While the occurrence of behaviour- and attitude-related content was independent of country, the problem-related content was specific to the U.S. context.

7.4.2. Consistency of probe responses

The second analytical focus was the consistency of probe responses to one another. Table 7.4 gives an overview of the consistency of probe responses as a function of question order. About one third (31%; *n* = 103) of probe responses were consistent to each other, with respondents either admitting to fare evasion ("I have done this on occasion") and displaying a lenient attitude (i.e., "It won't destroy the bus company") or denying having ever avoided the fare (i.e., "I always pay!") and showing a harsh attitude ("That's stealing"); 18% of responses were inconsistent (*n* = 60), with some respondents showing an awareness that their behaviour and attitude did not match. One respondent wrote in answer to the probe following the behaviour question: "That's about how many times I did this. Yeah, I'm a hypocrite and imperfect. I believe it's wrong to do this." The other 51% (*n* = 170) did not contain both behaviour- and attitude-related content; thus, these respondents were coded as unascertainable. Consistent probe responses were slightly more frequent in Germany (35%) than in the United States (27%); however, in both countries, about half of all probe responses were unascertainable.

Table 7.4. Consistency of probe responses, descriptive results

	Germany		**U.S.**	
	Behaviour question first	**Attitude question first**	**Behaviour question first**	**Attitude question first**
	% (n)	% (n)	% (n)	% (n)
Probe consistency				
consistent	33% (26)	38% (33)	29% (20)	25% (24)
inconsistent	14% (11)	15% (13)	19% (13)	24% (23)
unascertainable	54% (43)	47% (41)	53% (37)	51% (49)

Hypothesis 3 predicted that the consistency of probe responses is higher when the behaviour question is asked first. Hypothesis 4 predicted differences between countries regarding the consistency of probe responses. A multinomial logistic regression was carried out with consistent probe responses as the reference category. Main predictors were question order and country. Covariates were included as described under *Data Analysis*. Results are shown in Table 7.5 (Distributions of all predictors and covariates can be found in Table A.4 in the Appendix).

Table 7.5. Consistency of probe responses, multinomial logistic regression

Reference category: consistent probe responses	**Inconsistent**			**Unascertainable**		
	B	SE	*p*	B	SE	*p*
Question order (1=behaviour first)	0.19	0.34	0.58	-0.22	0.26	0.41
Country (1=Germany)	0.80*	0.39	0.04	0.17	0.29	0.56
Survey response to behaviour question (Q_beh; 1=admit)	-0.23	0.40	0.56	0.65*	0.31	0.03
Gender (1=women)	-0.21	0.35	0.55	0.53*	0.27	0.04
Age						
1=18-29 years	0.52	0.42	0.22	0.78*	0.33	0.02
2=30-49 years	0.73	0.41	0.07	0.95***	0.32	0.00
3=50-65 years	-	-	-	-	-	-
Education (1=low)	0.10	0.34	0.78	-0.31	0.26	0.24
Exaggeration of positive qualities	0.14	0.19	0.46	-0.02	0.15	0.92
Under exaggeration of negative qualities	-0.01	0.23	0.97	0.13	0.18	0.46
Constant	-1.73	0.99	0.08	-0.71	0.76	0.35
Model parameters	χ^2 = 34.332, df=18, *p* = .011, *N* = 333, $R^2_{(Nagelkerke)}$ = .113					

B = logit coefficient; SE = standard error. * *p* < .05; ** *p* < .01; *** *p* < .001.

Regarding the comparison of consistent to inconsistent probe responses, there was a significant effect of country (B = 0.80, SE = .39, *p* < .05). U.S. respondents were more likely to give inconsistent than consistent answers than German respondents. For the comparison of consistent to unascertainable probe responses, there were significant effects of the response to the behaviour survey question, gender, and age. Respondents admitting to the offense were more likely to give consistent and less likely to give unascertainable probe responses. Also, women and older respondents more likely to give consistent versus unascertainable probe responses. There was no significant effect of question order.

Hypothesis 3 could not be confirmed as there was no significant effect of question order on probe consistency. Hypothesis 4 could be confirmed, as there was a significant difference in probe consistency between countries, with German respondents more likely to give consistent responses than U.S. respondents.

7.5. Discussion and conclusion

The present research set out to explore whether and how question order impacts probe responses to behaviour and attitude questions in cross-cultural web probing. This was done by examining the impact of question order on the occurrence of probe content and the relation of the probe responses to one another in terms of self-reported attitude-behaviour consistency in two countries. Content and consistency of probe responses were not strongly impacted by question order in the present study.

There was limited support for the first hypothesis. Question order did not impact the occurrence of broad themes such as mentioning one's behaviour or attitude. The problem of the missing answer category for the behaviour question was more likely to be coded when this question was shown first. However, the case numbers for this model were rather low as the problem was restricted to the U.S. context. Further research is recommended to examine the occurrence of problem-related probe response content.

The second hypothesis predicted that respondents would be less likely to mention probe response content if they had already mentioned this content in answer to a previous probe. This hypothesis could not be confirmed. The occurrence of behaviour-related probe response content was irrespective of whether respondents had previously mentioned their behaviour. The occurrence of attitude-related content even increased when respondents had mentioned their attitude previously. Thus, there is no support that the norm of non-redundancy applies in the context of web probing. Possibly, asking several questions on the topic of fare evasion made respondents' related attitudes more salient (for saliency as a question order effect, see Bradburn, 1983). However, most respondents refrained from mentioning their attitude in response to the behaviour question and vice versa. More generally, the applicability of conversational norms in a self-administered web context must be questioned.

Contrary to the third hypothesis, the consistency of probe responses was not affected by question order. It was, however, significantly impacted by the response to the behaviour survey question. Respondents who did not admit to the behaviour were less likely to give substantive probe responses in the subsequent probing question, confirming previous results that the response to survey questions determines the likelihood of giving substantive probe responses (Zuell & Scholz, 2015).

In line with the fourth hypothesis, significant differences in probe response content and consistency were found between countries. The results demonstrated both differences in survey response behaviour between countries and differences that cannot be explained by country-specific survey response behaviour. There was no significant impact of country on the likelihood of including behaviour- or attitude-related content; however, problem-related probe content only emerged in the U.S. sample. Second, German respondents were more likely to demonstrate attitude-behaviour consistency in their probe responses, probably because German respondents were more likely to admit to fare evasion and align their attitude to be consistent with their self-reported behaviour. Finally, U.S. respondents were far less likely than German respondents to mention both their behaviour and attitude in one probe response. This result is in line with previous research that U.S. respondents generally offer less variety in themes (Meitinger et al., 2019). However, this finding can also be explained by the topic of public transport being less relevant to U.S. respondents. These findings underscore the importance of validating questionnaires in the language and country they are to be fielded in (see also Willis, 2015b). In summary, both predicting and explaining cross-cultural differences remain challenging, but including multiple countries in research designs is all the more crucial to make valid inferences about question order effects and to advance theorizing.

Finally, the results highlight that content-based coding makes the effects of question order and further contextual factors visible in ways that cannot be seen using standard measures of probe response quality, such as length of response or share of non-substantive answers. Future research in the area of web probing should combine quantitative and qualitative forms of analysis whenever possible.

The results are limited by several factors which light the way for future research. For one, the study examined question order effects by testing a behaviour and attitude question. This is certainly a highly relevant case for questionnaire designers and a common setup in cognitive pretesting. However, examining different types of survey questions may help uncover the underlying mechanisms of question order effects in pretesting. For instance, two behaviour or two attitude survey questions on the same topic might prove to be a better setting to examine the effects of redundancy and saliency.

Second, the current study employed only one probing technique. The use of general probes made it possible to use the same wording in both probing questions despite having different types of survey questions. However, other probing techniques, such as category selection probes or more specific probes, may lead to other effects (DeMaio & Landreth, 2004). Also, potential differences between concurrent probing, as employed in this study, and retrospective probing require further research.

Third, while web probing offers the ideal setup to quantify findings, the effects found must also be examined in the context of face-to-face cognitive interviewing in order to make inferences about question order effects on cognitive pretesting in general. The application of communication principles may be more pronounced in interviewer-administered modes.

Finally, the present study was carried out in two Western, post-industrial countries. Systematic research on cross-cultural differences in pretesting is desirable using a larger sample of countries (Meitinger et al., 2019; Pan, Landreth, Park, Hinsdale-Shouse, & Schoua-Glusberg, 2010; Park, Sha, & Pan, 2014) with different communicative principles (Grice, 1975).

Concluding, in the present study, probe response content and the consistency of probe responses were mostly not impacted by question order. This is good news regarding the stability and validity of cognitive pretest findings from web probing studies. At the same time, cross-cultural differences may impact the content and consistency of probe responses. In light of ever more complex survey and pretest designs, further methodological research is necessary to ensure that pretest results remain valid, reliable, and contribute to preventing measurement error in cross-cultural settings.

7.6. Data availability

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

7.7. Appendix Study 3

A.1. Questionnaire

Q_beh. As you know, many people occasionally commit minor offenses. We have listed four such minor offenses below. Please indicate for each of these behaviours, how often you have done this before. *[Wie Sie wissen, begehen viele Bürger hin und wieder eine kleinere Gesetzesübertretung. Im Folgenden sind vier solcher kleineren Gesetzesübertretungen genannt. Bitte kreuzen Sie bei jeder dieser vier Verhaltensweisen an, wie oft Sie in Ihrem Leben so etwas schon getan haben.]*

Item: Used public transport without having a valid ticket *[Öffentliche Verkehrsmittel benutzt, ohne dafür einen gültigen Fahrausweis zu besitzen]*

Response options: Never, Once, 2 to 5 times, 6 to 10 times, 11 to 20 times, More than 20 times, Don't know *[Noch nie, 1mal, 2 bis 5 mal, 6 bis 10 mal, 11 bis 20 mal, mehr als 20 mal, Kann ich nicht sagen]*

P_beh. We would like to know more about some of the previous statements and your answers. The question asked how often you have committed minor offenses. The first statement read: "[Q_beh]". Your answer was: [ANSWER]. What were you thinking of when you answered the question? *[Wir möchten gerne zu einigen der vorangegangenen Aussagen und Ihren Antworten noch nähere Informationen erhalten. In dieser Frage ging es darum, wie häufig man Gesetzesübertretungen begangen hat. Die erste Handlung lautete: „Öffentliche Verkehrsmittel benutzt, ohne dafür einen gültigen Fahrausweis zu besitzen“ Ihre Antwort lautete: [ANSWER]. Woran haben Sie beim Beantworten der Frage gedacht?]*

Q_att. Please indicate for each of the following actions whether you think it can always be justified, never be justified, or something in between. Please use the scale from 1 to 10, with 1 meaning an action "can never be justified", and 10 meaning that it "can always be justified". *[Bitte geben Sie für jede der folgenden Handlungen an, ob Sie sie in jedem Fall für in Ordnung halten, unter keinen Umständen für in Ordnung halten, oder irgendwas dazwischen. Bitte benutzen Sie die Skala von 1 bis 10, wobei 1 bedeutet:*

"Unter gar keinen Umständen in Ordnung" und 10 bedeutet: "In jedem Fall in Ordnung".]

Item: Avoiding a fare on public transport *[Kein Fahrgeld in öffentlichen Verkehrsmitteln zahlen, schwarzfahren]*

Response options: 1 never justifiable – 10 always justifiable *[1 Unter keinen Umständen – 10 In jedem Fall]*

P_att. We would like to know more about some of the previous statements and your answers. The question asked whether certain actions can always be justified (1), never be justified (10), or something in between. The second action read: "Avoiding a fare on public transport". Your answer was: [ANSWER]. What were you thinking of when you answered the question? *[Wir möchten gerne zu einigen der vorangegangenen Aussagen und Ihren Antworten noch nähere Informationen erhalten. In dieser Frage ging es darum, ob man bestimmte Handlungen in jedem Fall für in Ordnung hält (10), unter keinen Umständen für in Ordnung hält (1), oder irgendwas dazwischen. Die zweite Handlung lautete: „Kein Fahrgeld in öffentlichen Verkehrsmitteln zahlen, schwarzfahren". Ihre Antwort lautete: [ANSWER]. Woran haben Sie beim Beantworten der Frage gedacht?]*

A.2. Sample composition

	Behaviour question first	**Attitude question first**	χ^2	*p*
	N	*N*		
Germany	80	87		
U.S.	70	96		
Gender	% (*n*)	% (*n*)		
Germany				
Men	43.8% (35)	49.4% (43)	.539	.463
Women	56.3% (45)	50.6% (87)		
U.S.				
Men	37.1% (26)	45.8% (44)	1.254	.263
Women	62.9% (44)	54.2% (52)		
Age				
Germany				
18-29 years	35.0% (28)	29.9% (26)	.622	.733
30-49 years	28.8% (23)	33.3% (29)		
50-64 years	36.3% (29)	36.8% (32)		
U.S.				
18-29 years	27.1% (19)	33.3% (32)	1.005	.605
30-49 years	32.9% (23)	33.3% (32)		
50-64 years	40.0% (28)	33.3% (32)		
Education				
Germany				
High	53.8% (43)	54.0% (47)	.001	.972
Low	46.3% (37)	46.0% (40)		
U.S.				
High	50.0% (35)	50.0% (48)	.000	1.000
Low	50.0% (35)	50.0% (48)		

A.3. Cohen's Kappa for intercoder reliability

Probe	**Cohen's Kappa**
P_beh	
Behaviour-related content	0.855
Attitude-related content	0.850
P_att	
Behaviour-related content	0.916
Attitude-related content	0.905

A.4. Distributions of predictors and covariates of the multinomial logistic regression

		Probe Consistency			**Total**
		Consistent	**Inconsistent**	**Unascertainable**	
		n	*n*	*n*	*n*
Country	US	44	36	86	166
	Germany	59	24	84	167
Gender	Men	41	21	86	148
	Women	62	39	84	185
Age	18-29 years	28	19	58	105
	30-49 years	25	21	61	107
	50-64 years	50	20	51	121
Survey response to behaviour question	admit	41	23	48	112
	not admitted	62	37	122	221
Education	high	58	35	80	173
	low	45	25	90	160
Total		103	60	170	333

8. CONCLUSION

It is universally acknowledged that how questions are understood and answered is impacted by question context and that survey questions should be pretested and evaluated to ensure data quality. Employing probes to evaluate whether respondents understand an item as intended has become a cornerstone of cognitive pretesting methods, such as cognitive interviewing and web probing. Probing poses an introspection-based task in which respondents must report their thoughts during survey response. The mode of web probing additionally imposes the task of autonomously typing the response, as probes are usually administered in the form of open-ended questions. Naturally, the understanding of and response to probes is impacted by the context in which the probe is asked. The other way around, embedding probes into a survey changes the context in which survey questions are asked and answered. As a result, context effects that are the consequence of combining survey questions and probes are complex as they may impact both survey and probe responses. The potential directions of context effects depend on whether and in how far respondents can navigate independently through the survey.

In this thesis, I proposed a psychological model of context effects in web probing. The model differentiates between the three directions of effects: the effects of survey questions on probes, the effects of probes on survey questions and the effects of probes on each other. The model postulates that the mechanisms underlying these effects are response burden, the maxims of relation and quantity, reactivity, and memory errors. The model was tested in three empirical studies that sought to answer the following research questions:

Research Question 1: How do intermittent survey questions impact probe responses?

Research Question 2: How does embedding probes concurrently impact surrounding survey questions?

Research Question 3: How do probes pertaining to different survey questions with an overarching topic impact each other?

8.1. Summary of the results

The three studies differed in the dependent variables they examined, with studies 1 and 3 examining effects on probe responses and studies 2 and partially 3 effects on survey responses (see Figure 8.1). The first study examined whether retrospective probe placement—and thus intermittent questions between a survey question and the probe relating to it—increases the perceived response burden of answering probes and whether it promotes memory errors in probe response content. Probe placement increased the response burden in that respondents needed longer to read the probe and recapitulate the survey question. The share of non-substantive answers was increased for one of three probes, and respondents relied on memory cues from one of two topically related intermittent survey questions. The study also examined whether the adverse effects of retrospective probe placement can be decreased by employing probes with predefined response options. This was the case for relying on memory cues only; response latency and the share of non-substantive answers were increased through retrospective probe placement regardless of probe format.

The second study examined the impact of inserting concurrent web probes on the surrounding survey questions. Respondents were more likely to break off a survey when concurrent probes were asked, supporting the notion that probes increase the overall response burden of a survey (Luebker, 2021). Respondents were more likely to backtrack to previous survey pages and change their responses when presented with probes. This indicates that probing can cause respondents to re-evaluate their survey response and retroactively align their response with their given probe response, in line with the maxim of relation. Effects on subsequent survey questions occurred for two dependent measures for a question on relationship satisfaction only. The total response time taken to read and answer the survey question was longer, and respondents were more likely to give an extreme response (that is, report extreme satisfaction or dissatisfaction with their relationship) in the condition that included probes. This indicates that probing can cause reactivity by impacting how later survey questions are processed and answered. However, these effects only occurred for one of six tested questions, and the effect on response times was small. In contrast, an earlier study by Couper (2013) had demonstrated small but significant effects on the means of all items of a multi-item inventory when probes

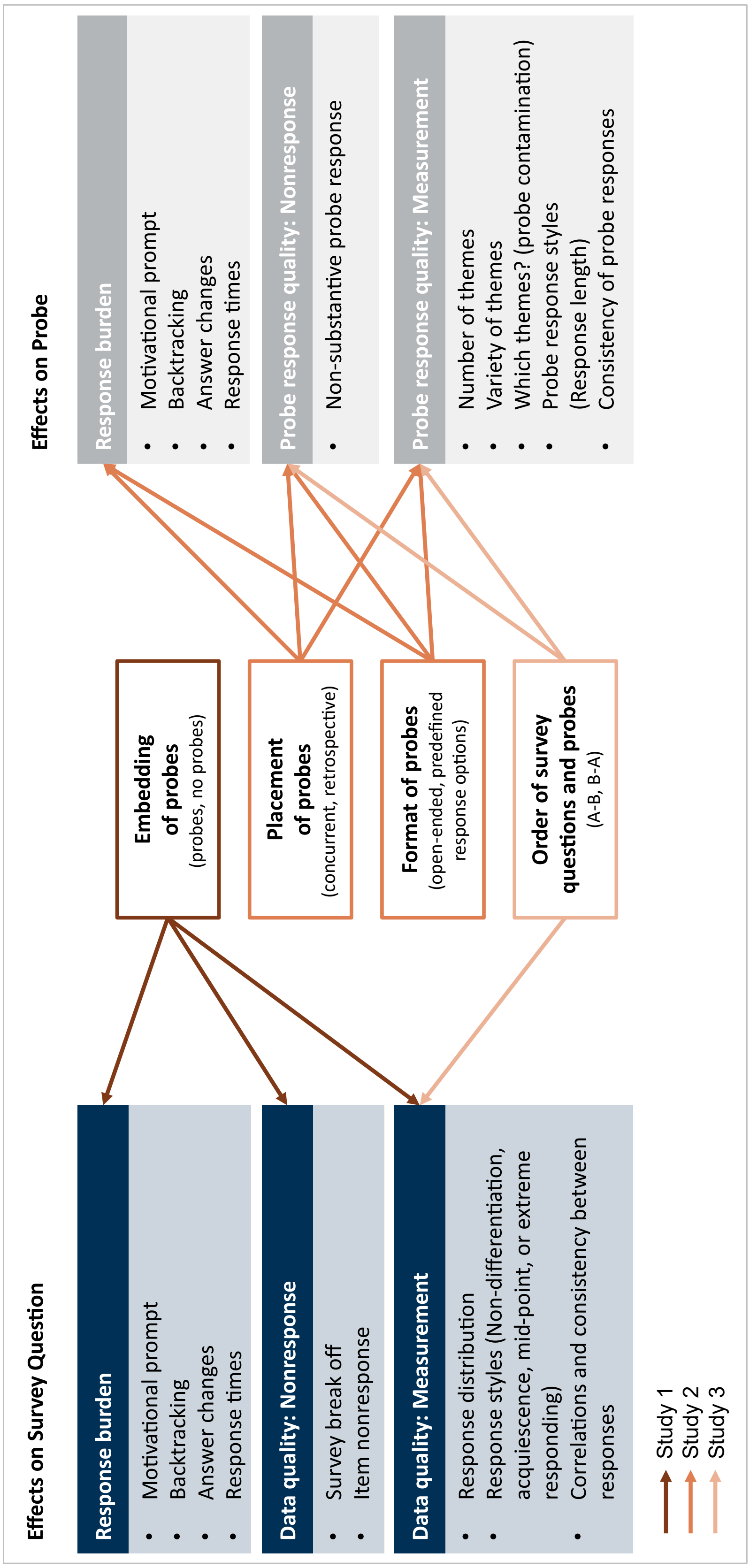

Effects on Survey Question
Response burden
• Motivational prompt
• Backtracking
• Answer changes
• Response times
Data quality: Nonresponse
• Survey break off
• Item nonresponse
Data quality: Measurement
• Response distribution
• Response styles (Non-differentiation, acquiescence, mid-point, or extreme responding)
• Correlations and consistency between responses
Study 1
Study 2
Study 3
Embedding of probes
(probes, no probes)
Placement of probes
(concurrent, retrospective)
Format of probes
(open-ended, predefined response options)
Order of survey questions and probes
(A-B, B-A)
Effects on Probe
Response burden
• Motivational prompt
• Backtracking
• Answer changes
• Response times
Probe response quality: Nonresponse
• Non-substantive probe response
Probe response quality: Measurement
• Number of themes
• Variety of themes
• Which themes? (probe contamination)
• Probe response styles
(Response length)
• Consistency of probe responses

Figure 8.1. Overview of empirical studies

were asked between the items. Potentially, the effects found in the present study only occurred for the question on relationship satisfaction because the preceding survey question was another measure of quality of life, namely life satisfaction. This is, however, a post hoc explanation and future research should clarify which factors contribute to probes impacting subsequent survey questions.

The third study examined the applicability of the maxims of relation and quantity to probe responses based on the order of survey questions and probes. To this end, one behavioural and one attitudinal survey question about the same delinquency were asked. A probe was placed directly following each survey question. The study included samples from Germany and the U.S. It was assumed that respondents would be less likely to mention behaviour-, attitude- or problem-related content in response to the second-shown probe, either because they had already mentioned the theme in response to the first-shown probe, or because they believed to have already offered all relevant information on the topic. Neither question order nor previously mentioning a topic decreased the likelihood of mentioning behaviour- or attitude-related content in response to the second-shown probe (for attitude-related content, previously mentioning one’s attitude even slightly increased the likelihood). However, many U.S. respondents encountered problem finding a suitable answer category to the behaviour survey question and were significantly less likely to report the problem when this question was shown second. The second analytical perspective focused on the relation of responses to one another. Probe responses related to and confirmed the chosen survey response. Some respondents used the probes to explain inconsistencies between their attitude and past behaviour, for instance, why they believed behaviour is not justifiable under any circumstances although they committed it themselves. Thus, respondents applied the maxim of relation in that their probe responses were coherent with the survey responses and each other. Based on the notion that attitude-behaviour consistency of survey questions is higher when the behaviour question was asked first, the study hypothesized that this would also be the case for probe responses. The study assumed that respondents would, for instance, be more likely to demonstrate a more lenient attitude towards the delinquency in their probe responses when they were asked the behaviour question first and admitted to the behaviour. This form of attitude-behaviour consistency of probe responses could not be demonstrated. The data showed that the choice of survey response to the behaviour question impacted the likelihood of providing substantive probe response content, with respondents who did not admit to the behaviour being significantly less likely to give a substantive probe response and respondents who admitted to having committed the delinquency being highly likely to explain and justify their survey response (in line with what Meitinger et al., 2022 have coined the need for justification). Thus, many probe responses could not be coded as consistent or inconsistent due to respondents not offering substantive content for the probe following the behaviour question. As the delinquent behaviour was much more commonplace in Germany, this led to cross-cultural differences in survey and subsequently probe response behaviour. In summary, there were no effects of question order on probe analyses following a descriptive approach, but an effect of question order on analysis following a reparative approach. Moreover, respondents answered the second-shown probe in light of the preceding survey questions and probe, though this was not measurable in terms of the attitude-behaviour consistency of the probe responses. From a methodological point of view, the operationalization of the maxims of relation and quantity may have been inadequate. The general codes of behaviour- and attitude-related content may be too broad to depict the effects of question order and previous mentions of themes on the later-shown probe. Moreover, the consistency of probe responses should not only have been gauged in terms of reported attitude-behaviour consistency but also in terms of coherence.

8.2. Implications

The findings of the empirical studies have implications for the psychological model of context effects in web probing established in Chapter 4, but also for web probing practice, the use of probes in cognitive interviews, and more generally for settings that combine open-ended and closed survey questions.

The studies confirmed the general applicability of context effects known from survey research to web probing through the mechanisms of response burden and communicative maxims; moreover, probing-specific effects of intro- and retrospection could be demonstrated through signs of reactivity and memory errors. Thus, overall, the model can be confirmed and may serve as a basis for further theorizing and empirical studies. An exception to this is the application of the maxim of quantity to probe response content, which requires more specification or even revision. On the one hand, past research has lent support for the norm on non-redundancy, with respondents expecting probes to require new information from them and voicing irritation when they believe this is not the case (Meitinger et al., 2022). At the same time, preceding probes seem to make some issues more salient (such as attitudinal content in study 3) leading to a repetition of at least broad themes. In contrast, other issues become less salient (such as problem-related content). The current model and the measures used to examine it cannot reconcile these findings. To advance the model and future research, more precise communicative theories and new measures of evaluating the relation of probe responses to one another and the relation between survey and probe responses are required.

In addition, the effects found in the empirical studies were often minor or only occurred in single cases. On the one hand, this is good news for researchers employing web probing. Probe responses are not genuinely volatile depending on aspects such as probe placement, the order of survey questions, or the presence of other probes. Likewise, survey data quality is not detrimentally impacted by embedding probes. On the other hand, that effects did not occur in all cases implies that the model deserves specification, so that researchers know when to expect which types of contextual effects in web probing.

Despite these limitations, several practical implications can be drawn for researchers considering employing web probing. For one, integrating probes into web surveys increases the perceived response burden of the survey and may lead to an increase in survey break-off. Probes invite respondents to re-evaluate their survey response and—if the web survey design permits backtracking to previous questions—they may retroactively change the answer to the question the probe pertains to. Moreover, in single cases, embedding probes may impact how subsequent survey questions are processed and answered. Researchers may try to avoid these effects by placing probes retrospectively and disabling the option to return to previous survey pages. Retrospective placement, however, increases the response burden the probes impose on respondents and decreases probe response quality. Finally, when several probes pertain to an overarching topic, previous probes and probe responses can impact later ones by decreasing the likelihood of mentioning problems with survey questions. In summary, researchers should be aware that the question is not *whether* context effects occur in web probing. As survey questions and probes are always part of a communicative context, context effects are inevitable. Instead, researchers should be aware of which settings induce or at least promote which effects.

In many cases, context effects must not negatively impact the insights gained through web probing, provided researchers bear in mind that probe analysis remains a *qualitative* and *exploratory* method of question evaluation. The sequence of survey and probing questions impacted the prevalence of probe response content in both studies that had probe response content as a dependent measure; however, the respective content was named by at least some respondents in all settings. Introspection-based methods are prone to several fallacies which result in imprecise measurement, and while context effects contribute to these errors, they are far from the only source. As long as researchers remain aware that the frequency of a qualitatively assigned code cannot be equated with the prevalence of that theme or problem, web probing remains a powerful method to gain insights on how respondents construe the pragmatic meaning of survey questions and whether they encounter problems with them.

This thesis examined context effects when using cognitive probes in the specific setting of web probing. Web probing is becoming an increasingly popular question evaluation method, particularly to pretest cross-national surveys (Behr et al., 2020) and web surveys in the same survey mode (Fowler & Willis, 2020). Moreover, web probing provided the ideal setting to quantify the effects of question sequence by using experimental designs. The web survey mode also made it possible to collect client-side paradata unobtrusively and examine measures such as response times, returning to previous survey pages, answer changes, and the activation of motivational prompts. However, although the model of context effects was established for web probing, many aspects can also be applied to cognitive interviewing or even other survey settings involving open-ended questions. Indeed, none of the underlying mechanisms in the model are restricted to the web context. On the contrary, the communicative maxims of relation and quantity (Grice, 1975), introspection-based methods of reporting on thought processes (Bröder, 2019) and the notion of response burden were developed in the context of personal communication or interviewer-administered surveys. However, context effects are not entirely devoid of mode (Doušak, 2017), so a discussion of how far the model can be applied to other settings is called for. In a nutshell, context effects should likewise occur through the fallacies of intro- and retrospection and communicative maxims for probes asked in cognitive interviews, and the effects of response burden and communicative maxims should underly context effects involving other open-ended questions.

Many of the effects that the empirical studies of this thesis examined have been documented in cognitive interviews, though they were usually not established using experimental designs or large case numbers. For instance, examples of respondents changing their survey answers after responding to probes during cognitive interviews are well documented in research reports (i.e., Hadler, Lenzner et al., 2022, p. 43; Hadler et al., 2017, p. 61). Also, that respondents are prone to memory errors and fill memory gaps with coherent memory cues or based on general knowledge has been documented in both cognitive and qualitative interviews (Bowers & Snyder, 1990; Daugherty et al., 2001; Kuusela & Paul, 2000; Nisbett & Wilson, 1977). That probing may cause reactivity and lead to respondents processing and answering subsequent questions differently is likewise included as advice in textbooks on cognitive interviewing (Collins, 2015). However, researchers should remember that cognitive interviews are a laboratory method, and participants may generally process and respond to survey questions differently than they would outside the laboratory. In contrast, web probing (of self-administered questionnaires) takes place in a setting similar to the later survey. Context effects in web probing and cognitive interviewing differ regarding the role of perceived response burden. Cognitive interviews benefit from the personal rapport between the interviewer and participant, and probe responses need only be given orally. In subsequence, asking multiple probes after a survey question is generally considered unproblematic in cognitive interviews, whereas it is employed sparingly in web probing (Behr et al., 2017). Thus, the role of response burden for context effects in cognitive interviews merits further research.

Lastly, some of the mechanisms of the model can be applied to other types of open-ended survey questions. Singer and Couper (2017) have argued for implementing open-ended questions into production surveys (outside of cognitive pretesting and other qualitative studies) for purposes such as encouraging more truthful answers or understanding reasons for item nonresponse. Based on previous research, the effects of response burden and question context apply to open-ended questions (i.e., Galesic, 2006; Peytchev, 2009; Yan & Williams, 2022), as does the maxim of relation (i.e., Silber et al., 2020). Introspection on thought processes during survey response is specific to probing. At the same time, the lower prevalence of any given theme in response to open-ended compared to closed questions (Reja et al., 2003; Schuman & Presser, 1979) may be attributed to memory errors. Furthermore, some open-ended questions, such as the question on the most important issue a country is facing (Schuman et al., 1986), may promote a more central route to processing (Kahneman, 2012; Petty & Cacioppo, 1986), potentially impacting the route to processing that respondents take to subsequent questions. Whether and in how such mechanisms apply to other open-ended questions should, however, be the subject of future research.

8.3. Suggestions for future research

The main goal of this thesis was to gain insights into context effects in web probing by establishing a model of effects and their underlying mechanisms and examining them in a series of empirical studies. The findings support many of the assumptions underlying the model. At the same time, the limitations of the studies suggest multiple directions for future research.

First, not all relevant settings could be examined in this thesis, so parts of the model still require examination. For instance, investigating the effects of the order of the survey questions and probe placement in conjunction would be a valuable extension of the three studies, as this setting could demonstrate simultaneous effects on survey and probe responses. Moreover, studies collecting client-side survey navigation or even eye-tracking data (Neuert, 2016; Romano Bergstrom & Schall, 2014) would be helpful to examine how respondents process probes that are embedded alongside survey questions on the same page, as embedded probe placement was not examined in any of the reported studies.

Secondly, future studies should examine the impact of implementing probes on the psychometric properties of survey data (Raykov & Marcoulides, 2011). For instance, the reliability of scales administered with and without probes should be compared using measures of internal consistency, such as composite and test-retest reliability (see Menold & Raykov, 2016; Wilson et al., 1996 for similar study designs). Previous research has used probes to explore reasons for the lack of measurement invariance (Leitgöb et al., 2022; Meitinger, 2017). However, MGCFA would also be a suitable method to test whether the survey data collected by questions accompanied by probes is comparable to survey data without probes. Validation studies could examine whether probing can be used to improve the accuracy of behaviour self-reports (Singer & Couper, 2017). To name just one possibility, comparing the consistency of measurements using GPS data with self-reports on mobility (either accompanied by probes or not) would be a possibility to evaluate the effects of probing on response accuracy.

Third, the model and the results of the reported studies should serve as a starting point to make more precise predictions regarding context effects in web probing. In particular, differences between the effects of various probing techniques (Foddy, 1998) merit a more detailed examination. It is conceivable that probes that require respondents to reconsider their survey response, such as category selection probing, are more likely to promote backtracking and answer changes to previous survey questions than other techniques. Another conceivable effect is that comprehension probes are more likely than other techniques to cement a specific question interpretation, with consequences for subsequent questions on the same or a related topic.

Fourth, the thesis focussed on context effects caused by the sequence of survey questions and probes. However, context effects include more facets than question order (Smyth et al., 2007). For instance, how a web survey including probes is framed may impact survey and probe responses. One possibility it to inform respondents about probes at the onset of a web survey on the welcome page, or to frame the survey itself as a pretest. This results in a debriefed setting comparable to cognitive interviews but may impact the response rate of the web survey. On the other extreme, probes may be inserted into a production survey without any prior information. This prevents potential adverse effects on the response rate but may irritate respondents when they encounter the first probe.

Finally, the present thesis has examined web probing in its currently dominant form of open-ended questions in which respondents must *read* survey questions and probes and *type* their responses autonomously. However, technological development already permits alternative forms of question presentation, such as the possibility to have questions read aloud by the survey software (i.e., Höhne, 2023; Lenzner & Höhne, 2022) or to give responses orally through audio recording (i.e., Gavras, Höhne, Blom, & Schoen, 2022; Revilla & Couper, 2019). The currently clear distinction between cognitive interviewing as an interviewer-administered oral form of collecting verbal reports and web probing as a self-administered form of collecting written data will thus presumably not uphold in future. Though the effects of intro- and retrospection should not depend on whether questions are presented and answered in written or oral form, the role of response burden and communicative maxims may require re-examination.

To conclude, surveys remain “one of the most commonly used methods in the social sciences to understand the way societies work and to test theories about behaviour” (Groves et al., 2011, p. 3). Continuous and timely question evaluation methods are essential to ensure that survey questions fulfil this purpose (Willis, 2020), and cognitive methods including probing will undoubtedly remain part of the mix (Tourangeau, Maitland, Steiger, & Yan, 2020). This thesis hopefully promotes our understanding that both survey questions and probes constitute a form of communication between the researcher and the respondent. Therefore, a vigilant eye towards the communicative context of the questions asked—be they questions or questions about questions—remains a prerequisite for high-quality data.

9. REFERENCES

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

Aichholzer, J., & Kritzinger, S. (2016). Kurzskala politischer Zynismus (KPZ). https://doi.org/10.6102/zis245

Ajzen, I. (1985). From intentions to actions: A theory of planned behavior. In J. Kuhl & J. Beckmann (Eds.), *Springer Series in Social Psychology. Action control: From cognition to behavior* (pp. 11–39). Springer.

Ajzen, I., & Fishbein, M. (1977). Attitude-behavior relations: A theoretical analysis and review of empirical research. *Psychological Bulletin*, *84*(5), 888–918. https://doi.org/10.1037/0033-2909.84.5.888

Ajzen, I., & Fishbein, M. (1980). *Understanding attitudes and predicting social behavior*. Englewood Cliffs, NJ: Prentice Hall.

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

ALLBUS (2000). *Allgemeine Bevölkerungsumfrage der Sozialwissenschaften. ZA3452*: https://www.gesis.org/en/allbus/contents-search/questionnaires/.

American Association for Public Opinion Research (AAPOR) (2016). *Standard definitions: Final dispositions of case codes and outcome rates for surveys* (9th edition): AAPOR.

Andrews, M. (2005). *Who is being heard? Response bias in open-ended responses in a large government employee survey*. AAPOR - ASA Section on Survey Research Methods, Miami Beach, Florida. Retrieved from http://www.asasrms.org/Proceedings/y2005/files/JSM2005-000924.pdf

Barabino, B., Salis, S., & Useli, B. (2014). Fare evasion in proof-of-payment transit systems: Deriving the optimum inspection level. *Transportation Research Part B: Methodological*, *70*(5), 1–17. https://doi.org/10.1016/j.trb.2014.08.001

Bauer, P. C., Barbera, P., Ackermann, K., & Venetz, A. (2017). Is the left-right scale a valid measure of ideology? Individual-level variation in associations with “left” and "right" and left-right self-placement. *Political Behavior*, *39*, 553–583. https://doi.org/10.1007/s11109-016-9368-2

Beatty, P. C. (2004). The dynamics of cognitive interviewing. In S. Presser, J. M. Rothgeb, M. P. Couper, J. T. Lessler, E. Martin, J. Martin,. . . C. Skinner (Eds.), *Methods for testing and evaluating survey questionnaires* (pp. 45–66). Hoboken, NJ: John Wiley & Sons. https://doi.org/10.1002/0471654728.ch3

Beatty, P. C., Collins, D., Kaye, L., Padilla, J.-L., Willis, G. B., & Wilmot, A. (Eds.) (2020). *Advances in questionnaire design, development, evaluation and testing*. Hoboken, NJ: John Wiley & Sons.

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., 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., Bandilla, W., Kaczmirek, L., Braun, M., & Majer, S. (2011). *Probing in web surveys and response burden: First results from an international web survey*. International Workshop on "Comparative Survey Design and Implementation" (CSDI). Retrieved from https://csdiworkshop.org/wp-content/uploads/2020/03/Behr_CSDI_2011_final_I.pdf

Behr, D., Braun, M., Kaczmirek, L., & Bandilla, W. (2012). Testing the validity of gender ideology items by implementing probing questions in web surveys. *Field Methods*, *25*(2), 124–141. https://doi.org/10.1177/1525822X12462525

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., Kaczmirek, L., Bandilla, W., & Braun, M. (2012a). Asking Probing Questions in Web Surveys. *Social Science Computer Review*, *30*(4), 487–498. https://doi.org/10.1177/0894439311435305

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/10.1177/0894439311435305

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. http://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, NJ: John Wiley & Sons.

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

Bem, D. J. (1967). Self-perception: An alternative interpretation of cognitive dissonance phenomena. *Psychological Review*, *74*(3), 183–200. https://doi.org/10.1037/h0024835

Benítez, I., & Padilla, J.-L. (2014). Analysis of nonequivalent assessments across different linguistic groups using a mixed methods approach. *Journal of Mixed Methods Research*, *8*(1), 52–68. https://doi.org/10.1177/1558689813488245

Benítez, I., van de Vijver, F., & Padilla, J.-L. (2022). A mixed methods approach to the analysis of bias in cross-cultural studies. *Sociological Methods & Research*, *51*(1), 237–270. https://doi.org/10.1177/0049124119852390

Benítez Baena, I., & Padilla, J.-L. (2014). Cognitive interviewing in mixed research. In K. Miller, S. Willson, V. Chepp, & J.-L. Padilla (Eds.), *Cognitive interviewing methodology* (pp. 133–152). Hoboken, NJ: John Wiley & Sons.

Bergstrom, J. C. R., Erdman, C., & Lakhe, S. (2016). Navigation buttons in web-based surveys: Respondents preferences revisited in the laboratory. *Survey Practice*, *9*(1), 1–10. https://doi.org/10.29115/SP-2016-0005

Bishop, G. F. (1992). Qualitative analysis of question-order and context effects: The use of think-aloud responses. In N. Schwarz & S. Sudman (Eds.), *Context effects in social and psychological research* (pp. 149–162). New York: Springer.

Bishop, G. F., Hippler, H.-J., Schwarz, N., & Strack, F. (1988). A comparison of response effects in self-administered and telephone surveys. In R. M. Groves, P. Biemer, L. E. Lyberg, J. T. Massey, W. L. Nicholls, & J. Waksberg (Eds.), *Telephone survey methodology* (pp. 321–340). New York: Wiley.

Bishop, G. F., Oldendick, R. W., & Tuchfarber, A. J. (1985). The importance of replicating a failure to replicate: Order effects on abortion items. *Public Opinion Quarterly*, *49*(1), 105. https://doi.org/10.1086/268904

Blair, J., & Conrad, F. G. (2011). Sample size for cognitive interview pretesting. *Public Opinion Quarterly*, *75*(4), 636–658. https://doi.org/10.1093/poq/nfr035

Bless, H., & Schwarz, N. (2010). Mental construal and the emergence of assimilation and contrast effects: The inclusion/exclusion model. In M. P. Zanna (Ed.), *Advances in experimental social psychology: Volume 42* (1st ed., pp. 319–373). Amsterdam: Elsevier.

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

Bradburn, N. M. (1978). Respondent burden. In L. G. Reeder (Ed.), *DHEW Publication No. PHS 79-3207. Health survey research methods* (pp. 49–54). Washington, DC: U.S. Government Printing Office. Retrieved from http://www.asasrms.org/Proceedings/papers/1978_007.pdf

Bradburn, N. M. (1983). Response effects. In P. Rossi, J. Wright, & A. Anderson (Eds.), *Handbook of Survey Research* (pp. 289–328). New York: Academic Press.

Bradburn, N. M., Sudman, S., & Wansink, B. (2004). *Asking questions - The definitive guide to questionnaire design: For market research, political polls, and social and health questionnaires*. *Research Methods for the Social Sciences*. Hoboken: John Wiley & Sons.

Braun, M. (2008). Using egalitarian items to measure men's and women's family roles. *Sex Roles*, *59*(9-10), 644–656. https://doi.org/10.1007/s11199-008-9468-5

Braun, M., Behr, D., & Díez Medrano, J. (2018). What do respondents mean when they report to be "citizens of the world"? Using probing questions to elucidate international differences in cosmopolitanism. *Quality & Quantity*, *52*(3), 1121–1135. https://doi.org/10.1007/s11135-017-0507-6

Braun, M., Behr, D., & Kaczmirek, L. (2013). Assessing cross-national equivalence of measures of xenophobia: Evidence from probing in web surveys. *International Journal of Public Opinion Research*, *25*(3), 383–395. https://doi.org/10.1093/ijpor/eds034

Braun, M., Behr, D., Kaczmirek, L., & Bandilla, W. (2014). Evaluating cross-national item equivalence with probing questions in web surveys. In U. Engel, B. Jann, P. Lynn, A. C. Scherpenzeel, & P. Sturgis (Eds.), *European Association of Methodology. Improving survey methods: Lessons from recent research* (pp. 184–200). New York, London: Routledge Taylor & Francis Group.

Braun, M., & Johnson, T. P. (2018). How should immigrants adapt to their country of residence? A mixed methods approach to evaluate the international applicability of a question from the German General Social Survey (ALLBUS). In E. Davidov, P. Schmidt, J. Billiet, & B. Meuleman (Eds.), *Cross-cultural analysis: methods and applications* (pp. 615–632). New York: Routledge.

Braun, M., Meitinger, K., & Behr, D. (2020). Combining quantitative experimental data with web probing: The case of individual solutions for the division of labor between both genders. *methods, data, analyses*, 1–25. https://doi.org/10.12758/mda.2020.03

Bredenkamp, J. (1990). Kognitionspsychologische Untersuchungen eines Rechenkünstlers. In H. Feger (Ed.), *Wissenschaft und Verantwortung* (pp. 47–70). Göttingen: Hogrefe.

Brentano, F. (2014 [1874]). *Psychology from an empirical standpoint*. *Routledge Classics*. Hoboken: Taylor and Francis.

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.

Bruijne, M. A. de (2015). *Designing web surveys for the multi-device internet*. *Doctoral dissertation*. Tilburg University: Tilburg, NL. Retrieved from https://pure.uvt.nl/ws/portalfiles/portal/8728830/Thesis_MarikadeBruijne.pdf

Bruin, A. de, Picavet, H. S. J., & Nossikov, A. (1996). *Health interview surveys: Towards international harmonization of methods and instruments*. *WHO Regional Publications, European Series: Vol. 58*. Retrieved from https://apps.who.int/iris/handle/10665/107328

Budd, R. J. (1987). Response bias and the theory of reasoned action. *Social Cognition*, *5*(2), 95–107. https://doi.org/10.1521/soco.1987.5.2.95

Buehler, R. (2011). Determinants of transport mode choice: A comparison of Germany and the USA. *Journal of Transport Geography*, *19*(4), 644–657. https://doi.org/10.1016/j.jtrangeo.2010.07.005

Buehler, R., & Pucher, J. (2012). Demand for public transport in Germany and the USA: An analysis of rider characteristics. *Transport Reviews*, *32*(5), 541–567. https://doi.org/10.1080/01441647.2012.707695

Bühler, K. (1907). Tatsachen und Probleme zu einer Psychologie der Denkvorgänge. I. Über Gedanken. *Archiv für die gesamte Psychologie*, *9*, 297–365.

Bühler, K. (1908). Antwort auf die von W. Wundt erhobenen Einwände gegen die Methode der Selbstbeobachtung an experimentell erzeugten Erlebnissen. *Archiv für die gesamte Psychologie*, *12*, 93–122.

Callegaro, M., & DiSogra, C. (2008). Computing response metrics for online panels. *Public Opinion Quarterly*, *72*(5), 1008–1032. https://doi.org/10.1093/poq/nfn065

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

Chang, L., & Krosnick, J. A. (2009). National surveys via Rdd telephone interviewing versus the internet: Comparing sample representativeness and response quality. *Public Opinion Quarterly*, *73*(4), 641–678. https://doi.org/10.1093/poq/nfp075

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

Chen, S., & Chaiken, S. (1999). The heuristic-systematic model in its broader context. In S. Chaiken & Y. Trope (Eds.), *Dual-process theories in social psychology* (pp. 73–96). New York, NY: Guilford Press.

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, NJ: 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.

Conrad, F. G., & Blair, J. (2009). Sources of error in cognitive interviews. *Public Opinion Quarterly*, *73*(1), 32–55. https://doi.org/10.1093/poq/nfp013

Conrad, F. G., Blair, J., & Tracy, E. (1999). *Verbal reports are data! A theoretical approach to cognitive interviews* (Proceedings of the 1999 Federal Committee on Statistical Methodology Research Conference). Washington, DC. Retrieved from https://www.bls.gov/osmr/research-papers/1999/st990240.htm

Conrad, F. G., Schober, M. F., & Schwarz, N. (2014). Pragmatic processes in survey interviewing. In T. M. Holtgraves (Ed.), *The Oxford handbook of language and social psychology* (pp. 420–437). Oxford: Oxford University Press.

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

Cools, M., Fabbro, Y., & Bellemans, T. (2018). Identification of the determinants of fare evasion. *Case Studies on Transport Policy*, *6*(3), 348–352. https://doi.org/10.1016/j.cstp.2017.10.007

Cornesse, C., & Blom, A. G. (2020). Response quality in nonprobability and probability-based online panels. *Sociological Methods & Research*, *21*(1), online first. https://doi.org/10.1177/0049124120914940

Couper, M. P. (2000). Web surveys: A review of issues and approaches. *Public Opinion Quarterly*, *64*(4), 464–494. https://doi.org/10.1086/318641

Couper, M. P. (2013). Research note: Reducing the threat of sensitive questions in online surveys? *Survey Methods: Insights from the Field.*, 1–9. https://doi.org/10.13094/SMIF-2013-00008

Couper, M. P., Baker, R., & Mechling, J. (2011). Placement and design of navigation buttons in web surveys. *Survey Practice*, *4*(1), 1–11. https://doi.org/10.29115/SP-2011-0001

Couper, M. P., Conrad, F. G., & Tourangeau, R. (2007). Visual context effects in web surveys. *Public Opinion Quarterly*, *71*(4), 623–634. https://doi.org/10.1093/poq/nfm044

Couper, M. P., Kennedy, C., Conrad, F. G., & Tourangeau, R. (2011). Designing input fields for non-narrative open-ended responses in web surveys. *Journal of Official Statistics*, *27*(1), 65–85.

Couper, M. P., & Kreuter, F. (2013). Using paradata to explore item level response times in surveys. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *176*(1), 271–286. https://doi.org/10.1111/j.1467-985X.2012.01041.x

Couper, M. P., Tourangeau, R., & Kenyon, K. (2004). Picture this! Exploring visual effects in web surveys. *Public Opinion Quarterly*, *68*(2), 255–266. https://doi.org/10.1093/poq/nfh013

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*. Proceedings of the annual meeting of the American Statistical Association. Retrieved from https://www.researchgate.net/profile/Lauren-Harris-Kojetin/publication/266866573_MAXIMIZING_THE_QUALITY_OF_COGNITIVE_INTERVIEWING_DATA_AN_EXPLORATION_OF_THREE_APPROACHES_AND_THEIR_INFORMATIONAL_CONTRIBUTIONS/links/54d108b90cf25ba0f0409c5a/MAXIMIZING-THE-QUALITY-OF-COGNITIVE-INTERVIEWING-DATA-AN-EXPLORATION-OF-THREE-APPROACHES-AND-THEIR-INFORMATIONAL-CONTRIBUTIONS.pdf

DeMaio, T., & Landreth, A. (2004). Do different cognitive interview techniques produce different results? In S. Presser, J. M. Rothgeb, M. P. Couper, J. T. Lessler, E. Martin, J. Martin,. . . C. Skinner (Eds.), *Methods for testing and evaluating survey questionnaires* (pp. 89–108). Hoboken, NJ: John Wiley & Sons.

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

Dillman, D. A., Smyth, J. D., & Christian, L. M. (2014). *Internet, phone, mail, and mixed-mode surveys: The tailored design method* (4th edition). New Jersey: Wiley.

Doušak, M. (2017). Survey mode as a moderator of context effects. *Advances in Methodology and Statistics*, *14*(1), 1–17. https://doi.org/10.51936/epsf8017

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). Comparing traditional and crowdsourcing methods for pretesting survey questions. *SAGE Open*, *6*(4), 1-14. https://doi.org/10.1177/2158244016671770

Edwards, L. M., & Lopez, S. J. (2006). Perceived family support, acculturation, and life satisfaction in Mexican American youth: A mixed-methods exploration. *Journal of Counseling Psychology*, *53*(3), 279–287. Retrieved from https://epublications.marquette.edu/edu_fac/47?utm_source=epublications.marquette.edu%2Fedu_fac%2F47&utm_medium=PDF&utm_campaign=PDFCoverPages

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 edition (Revised edition): MIT Press.

Evans, J. S. B. T. (2003). In two minds: Dual-process accounts of reasoning. *Trends in Cognitive Sciences*, *7*(10), 454–459. https://doi.org/10.1016/j.tics.2003.08.012

Fazio, R. H., & Zanna, M. P. (1981). Direct experience and attitude-behavior consistency. *Advances in Experimental Social Psychology*, *14*, 161–202. https://doi.org/10.1016/S0065-2601(08)60372-X

Fazio, R. H., Zanna, M. P., & Cooper, J. (1978). Direct experience and attitude-behavior consistency: An information processing analysis. *Personality and Social Psychology Bulletin*, *4*(1), 48–51. https://doi.org/10.1177/014616727800400109

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

Festinger, L. (1957). *A theory of cognitive dissonance*: Stanford University Press.

Fiedler, K., Ackerman, R., & Scarampi, C. (2019). Metacognition: Monitoring and controlling one's own knowledge, reasoning and decisions. In R. J. Sternberg & J. Funke (Eds.), *The psychology of human thought: An introduction* (pp. 89–111). Heidelberg: Heidelberg University Publishing.

Fitzgerald, R., Widdop, S., Gray, M., & Collins, D. (2011). Identifying sources of error in cross-national questionnaires: Application of an error source typology to cognitive interview data. *Journal of Official Statistics*, *27*(4), 569–599. Retrieved from https://openaccess.city.ac.uk/id/eprint/1160/

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

Forsyth, B., Rothgeb, J. M., & Willis, G. B. (2004). Does questionnaire pretesting make a difference? An empirical test. In S. Presser, J. M. Rothgeb, M. P. Couper, J. T. Lessler, E. Martin, J. Martin,. . . C. Skinner (Eds.), *Methods for testing and evaluating survey questionnaires* (pp. 525–546). Hoboken, NJ: John Wiley & Sons.

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 questionnaire design, development, evaluation and testing* (pp. 451–469). Hoboken, NJ: 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 procedures 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

Friborg, O., & Rosenvinge, J. H. (2013). A comparison of open-ended and closed questions in the prediction of mental health. *Quality & Quantity*, *47*(3), 1397–1411. https://doi.org/10.1007/s11135-011-9597-8

Fry, A., Mitchell, S. A., & Wiener, L. (2021). Considerations for conducting and reporting digitally supported cognitive interviews with children and adults. *Journal of Patient-Reported Outcomes*, *5*(131), 1–8. https://doi.org/10.1186/s41687-021-00371-5

Fuchs, D., & Klingemann, H.-D. (1989). Das Links-Rechts-Schema als politischer Code: ein interkultureller Vergleich auf inhaltsanalytischer Grundlage. In M. Haller, H.-J. Hoffmann-Nowotny, & W. Zapf (Eds.), *Kultur und Gesellschaft: Verhandlungen des 24. Deutschen Soziologentags, des 11. Österreichischen Soziologentags und des 8. Kongresses der Schweizerischen Gesellschaft für Soziologie in Zürich 1988* (Vol. 24, pp. 484–498). Frankfurt am Main: Campus-Verlag.

Fuchs, D., & Klingemann, H.-D. (1990). The left-right schema. In M. K. Jennings & J. W. van Deth (Eds.), *De Gruyter Studies on North America: Vol. 5. Continuities in political action: A longitudinal study of political orientations in three Western democracies* (pp. 203–234). Berlin, Boston: De Gruyter. https://doi.org/10.1515/9783110882193.203

Fürst, E., & Herold, D. (2018). Fare evasion and ticket forgery in public transport: Insights from Germany, Austria and Switzerland. *Societies*, *8*(4), 98. https://doi.org/10.3390/soc8040098

Galesic, M. (2006). Dropouts on the web: Effects of interest and burden experienced during an online survey. *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

Galesic, M., & Bosnjak, M. (2009). Effects of questionnaire length on participation and indicators of response quality in a web survey. *Public Opinion Quarterly*, *73*(2), 349–360. https://doi.org/10.1093/poq/nfp031

Galesic, M., & Tourangeau, R. (2007). What is sexual harassment? It depends on who asks! Framing effects on survey responses. *Applied Cognitive Psychology*, *21*(2), 189–202. https://doi.org/10.1002/acp.1336

Garbarski, D., Schaeffer, N. C., & Dykema, J. (2015). The effects of response option order and question order on self-rated health. *Quality of Life Research*, *24*(6), 1443–1453. https://doi.org/10.1007/s11136-014-0861-y

Gaskell, G. D., Wright, D. B., & O'Muircheartaigh, C. (1995). Context effects in the measurement of attitudes: A comparison of the consistency and framing explanations. *British Journal of Social Psychology*, *34*(4), 383–393. https://doi.org/10.1111/j.2044-8309.1995.tb01072.x

Gavras, K., & 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*, *37*(3), 1–7. https://doi.org/10.1080/13645579.2020.1860279

Gavras, K., Höhne, J. K., Blom, A. G., & 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 (Statistics in Society)*, *185*(3), 872–890. https://doi.org/10.1111/rssa.12807

Geer, J. G. (1991). Do open-ended questions measure "salient" issues? *Public Opinion Quarterly*, *55*(3), 360–370. https://doi.org/10.1086/269268

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

Gerlitz, J.-Y. (2014). Intergenerationale familiale Verpflichtung. Advance online publication. https://doi.org/10.6102/zis212

GESIS (2020). Allgemeine Bevölkerungsumfrage der Sozialwissenschaften ALLBUS - Kumulation 1980-2018. ZA5274 Datenfile Version 1.0.0. https://doi.org/10.4232/1.13395

GLES (2019). Pre-election Cross Section (GLES 2017). https://doi.org/10.4232/1.13234

Grice, H. P. (1975). Logic and conversation. In P. Cole & J. L. Morgen (Eds.), *Syntax and Semantics: Volume 3: Speech Acts* (5th ed., pp. 41–58). New York: Academic Press.

Groves, R. M. (2011). Three eras of survey research. *Public Opinion Quarterly*, *75*(5), 861–871. https://doi.org/10.1093/poq/nfr057

Groves, R. M., Fowler, F. J., Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2011). *Survey Methodology* (2nd ed.). Hoboken: John Wiley & Sons.

Groves, R. M., & Lyberg, L. E. (2010). Total survey error: Past, present, and future. *Public Opinion Quarterly*, *74*(5), 849–879. https://doi.org/10.1093/poq/nfq065

Groves, R. M., Singer, E., Corning, A. D., & Bowers, A. (1999). A laboratory approach to measuring the effects on survey participation of interview length, incentives, differential incentives, and refusal conversion. *Journal of Official Statistics*, *15*(2), 251–268. Retrieved from https://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/a-laboratory-approach-to-measuring-the-effects-on-survey-participation-of-interview-length-incentives-differential-incentives-and-refusal-conversion.pdf

Haberstroh, S., Oyserman, D., Schwarz, N., Kühnen, U., & Ji, L.-J. (2002). Is the interdependent self more sensitive to question context than the independent self? Self-construal and the observation of conversational norms. *Journal of Experimental Social Psychology*, *38*(3), 323–329. https://doi.org/10.1006/jesp.2001.1513

Hadler, P. (2021). Question order effects in cross-cultural web probing – Pretesting behavior and attitude questions. *Social Science Computer Review*, *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. *Sociological Methods & Research*, *27*(1), Online first. https://doi.org/10.1177/00491241231176846

Hadler, P., Lenzner, T., Schick, L., & Neuert, C. E. (2022). *European Working Conditions Survey 2024: Preparation and cognitive testing of the online questionnaire*. *Eurofound Working Paper: WPEF22035*. Retrieved from https://www.eurofound.europa.eu/sites/default/files/wpef22035.pdf

Hadler, P., Neuert, C. E., Lenzner, T., & Menold, N. (2018). European Working Conditions Survey (EWCS): Cognitive Pretest. *GESIS Project Report.* https://doi.org/10.17173/pretest72

Hadler, P., Neuert, C. E., Lenzner, T., Stiegler, A., Sarafoglou, A., Bous, P., . . . Menold, N. (2017). RESPOND - Improving regional health system responses to the challenges of migration through tailored interventions for asylum-seekers and refugees. Cognitive Pretest. *GESIS Project Report.* https://doi.org/10.17173/pretest83

Hadler, P., Neuert, C. E., Ortmanns, V., & Stiegler, A. (2022). Are you…? Asking questions on sex with a third category in Germany. *Field Methods*, *34*(2), 91–107. https://doi.org/10.1177/1525822X211072326

Head, B. F., Dean, E., Flanigan, T., Swicegood, J., & Keating, M. D. (2016). Advertising for cognitive interviews: A comparison of facebook, Craigslist, and snowball recruiting. *Social Science Computer Review*, *34*(3), 360–377. https://doi.org/10.1177/0894439315578240

Heerwegh, D. (2003). Explaining response latencies and changing answers using client-side paradata from a web survey. *Social Science Computer Review*, *21*(3), 360–373. https://doi.org/10.1177/0894439303253985

Heerwegh, D. (2011). Internet survey paradata. In M. Das, P. Ester, & L. Kaczmirek (Eds.), *European Association of Methodology Ser. Social and behavioral research and the internet: Advances in applied methods and research strategies* (pp. 325–348). New York: Routledge.

Heider, F. (1958). *The psychology of interpersonal relations*. Hoboken: John Wiley & Sons.

Höhne, J. K. (2023). Are respondents ready for audio and voice communication channels in online surveys? *International Journal of Social Research Methodology*, *26*(3), 335–342. https://doi.org/10.1080/13645579.2021.1987121

Höhne, J. K., & Schlosser, S. (2018). Investigating the adequacy of response time outlier definitions in computer-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). Investigating 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

Hoogendoorn, A. W. (2004). A questionnaire design for dependent interviewing that addresses the problem of cognitive satisficing. *Journal of Official Statistics*, *20*(2), 219.232. Retrieved from https://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/a-questionnaire-design-for-dependent-interviewing-that-addresses-the-problem-of-cognitive-satisficing.pdf

Huang, J. L., Curran, P. G., Keeney, J., Poposki, E. M., & DeShon, R. P. (2012). Detecting and deterring insufficient effort responding to surveys. *Journal of Business and Psychology*, *27*(1), 99–114. https://doi.org/10.1007/s10869-011-9231-8

Hyman, H. H., & Sheatsley, P. B. (1950). The current status of American public opinion. In J. C. Payne (Ed.), *The teaching of contemporary affairs; Twenty-first yearbook of the National Council for the Social Studies* (pp. 11–34). New York: National Education Association.

Irimata, K. E., & Scanlon, P. J. (2022). The Research and Development Survey (RANDS) during COVID-19. *Statistical Journal of the IAOS*, *38*(1), 13–21. https://doi.org/10.3233/SJI-210880

Jobe, J. B., & Mingay, D. J. (1990). Cognitive laboratory approach to designing questionnaires for surveys of the elderly. *Public Health Reports*, *105*(5), 518–524. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1580104/

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: 2017/1*. Mannheim, Germany: GESIS – Leibniz Institute for the Social Sciences. https://doi.org/10.21241/ssoar.51100

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

Kahneman, D. (2012). *Thinking, fast and slow*. London: Penguin Books.

Kemper, C. J., Beierlein, C., Bensch, D., Kovaleva, D., & Rammstedt, B. (2014). *Eine Kurzskala zur Erfassung des Gamma-Faktors sozial erwünschten Antwortverhaltens: Die Kurzskala Soziale Erwünschtheit-Gamma (KSE-G). GESIS Working Papers: 2012|25*. Köln: GESIS – Leibniz Institute for the Social Sciences. Retrieved from https://www.gesis.org/fileadmin/kurzskalen/working_papers/KSE_G_Workingpaper.pdf

Kim, Y., Dykema, J., Stevenson, J., Black, P., & Moberg, D. P. (2019). Straightlining: Overview of measurement, comparison of indicators, and effects in mail-web mixed-mode surveys. *Social Science Computer Review*, *37*(2), 214–233. https://doi.org/10.1177/0894439317752406

Knowles, E. S. (1988). Item context effects on personality scales: Measuring changes the measure. *Journal of Personality and Social Psychology*, *55*(2), 312–320. https://doi.org/10.1037/0022-3514.55.2.312

Knowles, E. S., Coker, M. C., Cook, D. A., Diercks, S. R., Irwin, M. E., Lundeen, E. J., . . . Sibicky, M. E. (1992). Order effects within personality measures. In N. Schwarz & S. Sudman (Eds.), *Context effects in social and psychological research* (pp. 221–236). New York: Springer. https://doi.org/10.1007/978-1-4612-2848-6_15

Kreuter, F. (Ed.) (2013). *Improving surveys with paradata: Analytic uses of process information*. Hoboken, New Jersey: John Wiley & Sons.

Kritzinger, S., Johann, D., Aichholzer, J., Glinitzer, K., Glantschnigg, C., Thomas, K., . . . Zeglovits, E. (2014). *AUTNES Rolling-Cross-Section Panel Study 2013: ZA5857 Data file Version 2.0.0*. Cologne.

Krosnick, J. A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology*, *5*(3), 213–236. https://doi.org/10.1002/acp.2350050305

Krosnick, J. A. (1999). Survey research. *Annual Review of Psychology*, *50*, 537–567. https://doi.org/10.1146/annurev.psych.50.1.537

Krosnick, J. A., & Smith, W. R. (1994). Attitude strength. In V. S. Ramachandran (Ed.), *Encyclopedia of human behavior* (pp. 279–289). San Diego, CA: Academic Press.

Kuhn, D., & Dean, J. D. (2010). Metacognition: A bridge between cognitive psychology and educational practice. *Theory Into Practice*, *43*(4), 268–273. https://doi.org/10.1207/s15430421tip4304_4

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

Kunz, T., & Hadler, P. (2020). *Web paradata in survey research*. *GESIS Survey Guidelines*. Mannheim, Germany: 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., Quoß, F., & Gummer, T. (2020). Using placeholder text in narrative open-ended questions in web surveys. *Journal of Survey Statistics and Methodology*. (Online First), 1–21. https://doi.org/10.1093/jssam/smaa039

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., 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

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

Leeuw, E. de (2018). Mixed-mode: Past, present, and future. *Survey Research Methods*, *12*(2), 75–89. https://doi.org/10.18148/srm/2018.v12i2.7402

Legewie, J., Gerlitz, J.-Y., Mühleck, K., Scheller, P., & Schrenker, M. (2007). *Dokumentation des International Social Justice Projekt 2006 für Deutschland*. *ISJP Arbeitsbericht: Vol. 118*. Berlin: Humboldt Universität zu Berlin.

Leitgöb, H., Seddig, D., Asparouhov, T., Behr, D., Davidov, E., Roover, K. de, . . . van de Schoot, R. (2022). Measurement invariance in the social sciences: Historical development, methodological challenges, state of the art, and future perspectives. *Social Science Research*, *5*(2), 1–30. https://doi.org/10.1016/j.ssresearch.2022.102805

Lenzner, T. (2011). *A psycholinguistic look at survey question design and response quality*. *Dissertation*. Mannheim: MADOC. Retrieved from https://madoc.bib.uni-mannheim.de/29478

Lenzner, T., Hadler, P., & Neuert, C. E. (2022). An experimental test of the effectiveness of cognitive interviewing in pretesting questionnaires. *Quality & Quantity*, *16*(2), 296. https://doi.org/10.1007/s11135-022-01489-4

Lenzner, T., & Höhne, J. K. (2022). Who is willing to use audio and voice inputs in smartphone surveys, and why? *International Journal of Market Research*, *9*(2), 1-17. https://doi.org/10.1177/14707853221084213

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. (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., 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

Leys, C., Ley, C., Klein, O., Bernard, P., & Licata, L. (2013). Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. *Journal of Experimental Social Psychology*, *49*(4), 764–766. https://doi.org/10.1016/j.jesp.2013.03.013

Liska, A. E. (1984). A critical examination of the causal structure of the Fishbein/Ajzen attitude-behavior model. *Social Psychology Quarterly*, *47*(1), 61. https://doi.org/10.2307/3033889

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

Maitland, A., & Presser, S. (2016). How accurately do different evaluation methods predict the reliability of survey questions? *Journal of Survey Statistics and Methodology*, *4*(3), 362–381. https://doi.org/10.1093/jssam/smw014

Maitland, A., & Presser, S. (2018). How do question evaluation methods compare in predicting problems observed in typical survey conditions? *Journal of Survey Statistics and Methodology*, *6*(4), 465–490. https://doi.org/10.1093/jssam/smx036

Marcel, A. (2003). Introspective report: Trust, self-knowledge and science. *Journal of Consciousness Studies*, *10*(9-10), 167–186.

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

Meadows, K. (2021). Cognitive interviewing methodologies. *Clinical Nursing Research*, *30*(4), 375–379. https://doi.org/10.1177/1054773821101409

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. (2018). What does the general national pride item measure? Insights from web probing. *International Journal of Comparative Sociology*, *59*(5-6), 428–450. https://doi.org/10.1177/0020715218805793

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., 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., 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., & Kunz, T. (2022). Visual design and cognition in list-style open-ended questions in web probing. *Sociological Methods & Research*, *28*(1), 1-28. https://doi.org/10.1177/00491241221077241

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

Menold, N., Hadler, P., & Neuert, C. E. (2023). Improving cross-cultural comparability of measures on gender and age stereotypes by means of piloting methods. Advance online publication. https://doi.org/10.31124/advance.21716483.v1

Menold, N., & Raykov, T. (2016). Can reliability of multiple component measuring instruments depend on response option presentation mode? *Educational and Psychological Measurement*, *76*(3), 454–469. https://doi.org/10.1177/0013164415593602

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. (2014). Introduction. In K. Miller, S. Willson, V. Chepp, & J.-L. Padilla (Eds.), *Cognitive interviewing methodology* (pp. 1–5). Hoboken, NJ: John Wiley & Sons.

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

Mockovak, W., & Kaplan, R. (2015). *Comparing results from telephone reinterview with unmoderated, online cognitive interviewing.* American Association for Public Opinion Research (AAPOR), Boston, MA.

Mohorko, A., & Hlebec, V. (2016). Degree of cognitive interviewer involvement in questionnaire pretesting on trending survey modes. *Computers in Human Behavior*, *62*, 79–89. https://doi.org/10.1016/j.chb.2016.03.021

Murphy, J., Keating, M. D., & Edgar, J. (2013). *Crowdsourcing in the cognitive interviewing process*. Proceedings of the 2013 Federal Committee on Statistical Methodology (FCSM) Research Conference. Retrieved from http://dc-aapor.org/2014%20conference%20slides/EdgarMurphyKeating.pdf

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

Narayan, S., & Krosnick, J. A. (1996). Education moderates some response effects in attitude measurement. *Public Opinion Quarterly*, *60*(1), 58. https://doi.org/10.1086/297739

Neuert, C. E. (2016). *Eye tracking in questionnaire pretesting*. Mannheim: MADOC. Retrieved from https://madoc.bib.uni-mannheim.de/40829

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

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

Neuert, C. E., Meitinger, K., Behr, D., & Schonlau, M. (2021). Editorial: The use of open-ended questions in surveys. *methods, data, analyses (Special Issue)*, *15*(1), 3–6. Retrieved from https://nbn-resolving.org/urn:nbn:de:0168-ssoar-73172-3

Neuert, C. E., Roßmann, J., & Silber, H. (2023). Using eye-tracking methodology to study grid question designs in web surveys. *Journal of Official Statistics*, *39*(1), 79–101. https://doi.org/10.2478/jos-2023-0004

Nießen, D., Partsch, M. V., Kemper, C. J., & Rammstedt, B. (2019). An English-language adaptation of the Social Desirability–Gamma Short Scale (KSE-G). *Measurement Instruments for the Social Sciences*, *1*(2), 1–10. https://doi.org/10.1186/s42409-018-0005-1

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

Noel, H. (2013). *Conducting cognitive interviews over the phone: Benefits and challenges*. American Association of Public Opinion Research, Boston, MA. Retrieved from http://www.asasrms.org/Proceedings/y2013/files/400282_500777.pdf

Oksenberg, L., Cannell, C. F., & Kalton, G. (1991). New strategies for pretesting survey questions. *Journal of Official Statistics*, *7*(3), 349–365. Retrieved from https://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/new-strategies-for-pretesting-survey-questions.pdf

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

Padilla, J.-L., & Benítez Baena, I. (2014). Validity evidence based on response processes. *Psicothema*, *26*(1), 136–144. https://doi.org/10.7334/psicothema2013.259

Padilla, J.-L., & Leighton, J. P. (2017). Cognitive interviewing and think aloud methods. In B. D. Zumbo & A. M. Hubley (Eds.), *Social Indicators Research Series (SINS, volume 69). Understanding and investigating response processes in validation research* (pp. 211–228). Cham: Springer. https://doi.org/10.1007/978-3-319-56129-5_12

Pan, Y., Landreth, A., Park, H., Hinsdale-Shouse, M., & Schoua-Glusberg, A. (2010). Cognitive interviewing in non-English languages: A cross-cultural perspective. In J. A. Harkness, M. Braun, B. Edwards, T. P. Johnson, L. E. Lyberg, P. P. Mohler,. . . T. W. Smith (Eds.), *Wiley series in survey methodology. Survey methods in multinational, multiregional and multicultural contexts* (pp. 91–113). Hoboken, NJ: John Wiley & Sons.

Park, H., Sha, M. M., & Pan, Y. (2014). Investigating validity and effectiveness of cognitive interviewing as a pretesting method for non-English questionnaires: Findings from Korean cognitive interviews. *International Journal of Social Research Methodology*, *17*(6), 643–658. https://doi.org/10.1080/13645579.2013.823002

Paulhus, D. L. (2002). Socially desirable responding. The evolution of a construct. In H. I. Braun, D. N. Jackson, & D. E. Wiley (Eds.), *The role of constructs in psychological and educational measurement* (pp. 49–69). Mahwah, NJ: Erlbaum.

Petty, R. E., & Cacioppo, J. T. (1986). *Communication and Persuasion: Central and Peripheral Routes to Attitude Change*. *Springer Series in Social Psychology*. New York, NY: Springer.

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

Peytchev, A., Couper, M. P., McCabe, S. E., & Crawford, S. D. (2006). Web survey design: Paging versus scrolling. *Public Opinion Quarterly*, *70*(4), 596–607. https://doi.org/10.1093/poq/nfl028

Pongratz, L. J. (1997). Die Kontroverse zwischen Wilhelm Wundt (1832–1920) und Karl Bühler (1879–1963). Analyse einer Wende der Psychologie. *Brentano-Studien*, *7*, 255–266.

Presser, S., & Blair, J. (1994). Survey pretesting: Do different methods produce different results? *Sociological Methodology*, *24*, 73–104. https://doi.org/10.2307/270979

Presser, S., Rothgeb, J. M., Couper, M. P., Lessler, J. T., Martin, E., Martin, J.,. . . Skinner, C. (Eds.) (2004). *Methods for testing and evaluating survey questionnaires*. Hoboken, NJ: 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

Priede, C., Jokinen, A., Ruuskanen, E., & Farrall, S. (2014). Which probes are most useful when undertaking cognitive interviews? *International Journal of Social Research Methodology*, *17*(5), 559–568. https://doi.org/10.1080/13645579.2013.799795

Prochazka, F. (2020). *Vertrauen in Journalismus unter Online-Bedingungen: Zum Einfluss von Personenmerkmalen, Qualitätswahrnehmungen und Nachrichtennutzung*. Wiesbaden: Springer VS.

Rammstedt, B., Beierlein, C., Brähler, E., Eid, M., Hartig, J., Kersting, M., . . . Weichselgartner, E. (2015). *Quality Standards for the Development, Application, and Evaluation of Measurement Instruments in Social Science Survey Research*. *RatSWD Working Paper Series: Vol. 245*. Berlin: SCIVERO Verlag. Retrieved from https://www.konsortswd.de/wp-content/uploads/RatSWD_WP_245.pdf

Rasinski, K. A., Lee, L., & Krishnamurty, P. (2012). Question order effects. In H. Cooper, P. M. Camic, D. L. Long, A. T. Panter, D. Rindskopf, & K. J. Sher (Eds.), *APA handbook of research methods in psychology, Volume 1. Foundations, planning, measures, and psychometrics* (pp. 229–248). Washington: American Psychological Association. https://doi.org/10.1037/13619-014

Raykov, T., & Marcoulides, G. A. (2011). *Introduction to psychometric theory*. New York: Routledge.

Read, B. (2019). Respondent burden in a mobile app: Evidence from a shopping receipt scanning study. *Survey Research Methods*, *13*(1), 45–71. https://doi.org/10.18148/srm/2019.v1i1.7379

Reips, U.-D. (2002). Context effects in web surveys. In B. Batinic, U.-D. Reips, & M. Bosnjak (Eds.), *Online social sciences* (pp. 69–80). Seattle, Washington: Hogrefe & Huber.

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

Revilla, M., & Couper, M. P. (2019). Improving the use of voice recording in a smartphone survey. *Social Science Computer Review*, *39*(6), 1056-1312. https://doi.org/10.1177/0894439319888708

Rodax, N., & Benetka, G. (2021). Debating experimental psychology's frontiers: Re-discovering Wilhelm Wundt's contribution to contemporary psychological research. *Human Arenas*, *4*(1), 48–63. https://doi.org/10.1007/s42087-020-00173-z

Rogelberg, S. G., Fisher, G. G., Maynard, D. C., Hakel, M. D., & Horvath, M. (2001). Attitudes toward surveys: Development of a measure and its relationship to respondent behavior. *Organizational Research Methods*, *4*(1), 3–25. https://doi.org/10.1177/109442810141001

Romano Bergstrom, J., & Schall, A. (2014). *Eye tracking in user experience design*. Amsterdam: Morgan Kaufmann/Elsevier.

Roßmann, J., Blumenstiel, J. E., & Steinbrecher, M. (2015). Why do respondents break off web surveys and does it matter? Results from four follow-up surveys. *International Journal of Public Opinion Research*, *27*(2), 289–302. https://doi.org/10.1093/ijpor/edv030

Rothgeb, J. M., Willis, G. B., & Forsyth, B. (2007). Questionnaire pretesting methods: Do different techniques and different organizations produce similar results? *Bulletin de Méthodologie Sociologique*, *96*(1), 5–31. https://doi.org/10.1177/075910630709600103

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. (2016). *Using targeted embedded probes to quantify cognitive interviewing findings*. International Conference on Questionnaire Design, Development, Evaluation, and Testing, Miami, Florida. Retrieved from https://ww2.amstat.org/meetings/qdet2/OnlineProgram/Program.cfm?date=11-11-16

Scanlon, P. J. (2018). *Cognitive evaluation of the National Center for Health Statistics' 2018 Research and Development Survey*. Hyattsville, MD: National Center for Health Statistics. Retrieved from https://wwwn.cdc.gov/qbank/report/Scanlon_NCHS_2018_RANDS3.pdf

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, NJ: John Wiley & Sons.

Schaeffer, N. C., & Dykema, J. (2020). Advances in the science of asking questions. *Annual Review of Sociology*, *46*(1), 37–60. https://doi.org/10.1146/annurev-soc-121919-054544

Schick, L., Lenzner, T., Hadler, P., & Neuert, C. E. (2023). *FReDA-W3b – Fragen zu den Themen Partnerschaftsstatus, Ernährungsstile, globale Unsicherheit und Vertrauen in Institutionen. Kognitiver Online-Pretest*. *GESIS Projektbericht*: GESIS – Leibniz Institute for the Social Sciences. https://doi.org/10.17173/pretest127

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, NY: John Wiley & Sons.

Schober, M. F., Conrad, F. G., Hupp, A. L., Larsen, K. M., Ong, A. R., & West, B. T. (2020). Design considerations for live video survey interviews. *Survey Practice*, *13*(1), 1–11. https://doi.org/10.29115/SP-2020-0014

Scholz, E., & Zuell, C. (2012). Item non-response in open-ended questions: Who does not answer on the meaning of left and right? *Social Science Research*, *41*(6), 1415–1428. https://doi.org/10.1016/j.ssresearch.2012.07.006

Schonlau, M., & Couper, M. P. (2016). Semi-automated categorization of open-ended questions. *Survey Research Methods*, *10*(2), 143–152. https://doi.org/10.18148/srm/2016.v10i2.6213

Schooler, J. W., Ohlsson, S., & Brooks, K. (1993). Thoughts beyond words: When language overshadows insight. *Journal of Experimental Psychology: General*, *122*(2), 166–183. https://doi.org/10.1037/0096-3445.122.2.166

Schul, Y., & Schiff, M. (1993). Measuring satisfaction with organizations. Predictions from information accessibility. *Public Opinion Quarterly*, *57*(4), 536–551. https://doi.org/10.1086/269394

Schuman, H. (1966). The random probe: A technique for evaluating the validity of closed questions. *American Sociological Review*, 218–222. https://doi.org/10.2307/2090907

Schuman, H., & Ludwig, J. (1983). The norm of even-handedness in surveys as in life. *American Sociological Review*, *48*(1), 112–120. https://doi.org/10.2307/2095149

Schuman, H., Ludwig, J., & Krosnick, J. A. (1986). The perceived threat of nuclear war, salience, and open questions. *Public Opinion Quarterly*, *50*(4), 519–536. https://doi.org/10.1086/269001

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

Schuman, H., & Presser, S. (1981). *Questions and answers in attitude surveys: Experiments on question form, wording, and context. (Quantitative studies in social relations)*. New York: Academic Press.

Schwarz, N. (1995). What respondents learn from questionnaires: The survey interview and the logic of conversation. *International Statistical Review / Revue Internationale de Statistique*, *63*(2), 153–168. https://doi.org/10.2307/1403610

Schwarz, N. (1996). *Cognition and communication: Judgmental biases, research methods, and the logic of conversation*. New York: Psychology Press.

Schwarz, N., & Bless, H. (1992). Constructing reality and its alternatives: Assimilation and contrast effects in social judgment. In L. L. Martin & A. Tesser (Eds.), *The construction of social judgments* (pp. 217–245). Hilsdale, NJ: Erlbaum.

Schwarz, N., & Bless, H. (2007). Mental construal processes: The inclusion/exclusion model. In D. A. Stapel & J. Suls (Eds.), *Assimilation and contrast in social psychology* (pp. 119–141). New York, NY: Psychology Press.

Schwarz, N., & Bohner, G. (2001). The construction of attitudes. In A. Tesser & N. Schwarz (Eds.), *Blackwell handbook of social psychology: / series eds.: Miles Hewstone. Intraindividual processes* (pp. 436–457). Malden, Mass.: Blackwell.

Schwarz, N., & Hippler, H.-J. (1995). Subsequent questions may influence answers to preceding questions in mail surveys. *Public Opinion Quarterly*, *59*(1), 93–97. https://doi.org/10.1086/269460

Schwarz, N., Hippler, H.-J., Deutsch, B., & Strack, F. (1985). Response scales: Effects of category range on reported behavior and comparative judgments. *Public Opinion Quarterly*, *49*(3), 388. https://doi.org/10.1086/268936

Schwarz, N., Knäuper, B., Oyserman, D., & Stich, C. (2008). The psychology of asking questions. In E. de Leeuw, J. J. Hox, & D. A. Dillman (Eds.), *EAM book series. International handbook of survey methodology* (pp. 18–34). New York: Psychology Press.

Schwarz, N., Oyserman, D., & Peytcheva, E. (2010). Cognition, communication, and culture: Implications for the survey response process. In J. A. Harkness, M. Braun, B. Edwards, T. P. Johnson, L. E. Lyberg, P. P. Mohler,. . . T. W. Smith (Eds.), *Wiley* *series in survey methodology*. *Survey methods in multinational, multiregional and multicultural contexts* (pp. 177–190). Hoboken, NJ: John Wiley & Sons.

Schwarz, N., & Strack, F. (1991). Context effects in attitude surveys: Applying cognitive theory to social research. *European Review of Social Psychology*, *2*(1), 31–50. https://doi.org/10.1080/14792779143000015

Schwarz, N., & Strack, F. (1999). Reports of subjective well-being: Judgmental processes and their methodological implications. In D. Kahneman, E. Diener, & N. Schwarz (Eds.), *Well-being: The foundations of hedonic psychology* (pp. 61–84). New York: Russell Sage Foundation.

Schwarz, N., Strack, F., Hippler, H.-J., & Bishop, G. F. (1991). The impact of administration mode on response effects in survey measurement. *Applied Cognitive Psychology*, *5*(3), 193–212. https://doi.org/10.1002/acp.2350050304

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

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., & Sudman, S. (Eds.) (1992). *Context effects in social and psychological research*. New York: Springer.

Schwarz, N., & Sudman, S. (Eds.) (1996). *Answering questions: Methodology for determining cognitive and communicative processes in survey research*. San Fransisco: Jossey-Bass.

Sharp, L. M., & Frankel, J. (1983). Respondent burden: A test of some common assumptions. *Public Opinion Quarterly*, *47*(1), 36–53. https://doi.org/10.1086/268765

Shoemaker, P. J., Eichholz, M., & Skewes, E. A. (2002). Item nonresponse: Distinguishing between don't know and refuse. *International Journal of Public Opinion Research*, *14*(2), 193–201. https://doi.org/10.1093/ijpor/14.2.193

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 (European Journal of Research Methods for the Behavioral and Social Sciences)*, *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

Sirken, M. G., Herrmann, D. J., Schechter, S., Schwarz, N., Tanur, J. M., & Tourangeau, R. (Eds.) (1999). *Cognition and survey research*. *Wiley series in probability and statistics Survey methodology section*. New York, NY: John Wiley & Sons.

Smith, T. W. (1982). *Conditional order effects*. GSS Technical Report no. 13. Chicago: NORC. Retrieved from https://gss.norc.org/Documents/reports/methodological-reports/MR020.pdf

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

Smith, T. W. (1992). Thoughts on the nature of context effects. In N. Schwarz & S. Sudman (Eds.), *Context effects in social and psychological research* (pp. 163–186). New York: Springer.

Smyth, J. D., Dillman, D. A., & Christian, L. M. (2007). Context effects in Internet surveys: New issues and evidence. In K. Y. A. McKenna, T. Postmes, U.-D. Reips, & A. N. Joinson (Eds.), *Oxford handbook of internet psychology* (Vol. 1, pp. 429–446). New York, NY: Oxford University Press.

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

Snijkers, G. (2002). *Cognitive laboratory experiences on pre-testing computerised questionnaires and data quality*. Statistics Netherlands: Heerlen. Retrieved from https://dspace.library.uu.nl/handle/1874/13401

Stark, T. H., Silber, H., Krosnick, J. A., Blom, A. G., Aoyagi, M., Belchior, A., . . . Yu, R.-r. (2020). Generalization of classic question order effects across cultures. *Sociological Methods & Research*, *49*(3), 567–602. https://doi.org/10.1177/0049124117747304

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)27:3<313::AID-EJSP824>3.0.CO;2-4

Strack, F. (1992). “Order effects” in survey research: Activation and information functions of preceding questions. In N. Schwarz & S. Sudman (Eds.), *Context effects in social and psychological research* (pp. 23–34). New York: Springer.

Strack, F., Schwarz, N., & Wänke, M. (1991). Semantic and pragmatic aspects of context effects in social and psychological research. *Social Cognition*, *9*(1), 111–125. https://doi.org/10.1521/soco.1991.9.1.111

Sudman, S., Bradburn, N. M., & Schwarz, N. (1996). *Thinking about answers: The application of cognitive processes to survey methodology* (1. ed.). San Francisco, CA: Jossey-Bass.

Tanur, J. M. (Ed.) (1992). *Questions about questions: Inquiries into the cognitive bases of surveys*: Russell Sage Foundation.

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

Toepoel, V., & Couper, M. P. (2011). Can verbal instructions counteract visual context effects in web surveys? *Public Opinion Quarterly*, *75*(1), 1–18. https://doi.org/10.1093/poq/nfq044

Toepoel, V., & Dillman, D. A. (2011). How visual design affects the interpretability of survey questions. In M. Das, P. Ester, & L. Kaczmirek (Eds.), *European Association of Methodology Ser. Social and behavioral research and the internet: Advances in applied methods and research strategies* (pp. 165–190). New York: Routledge.

Tourangeau, R. (1999). Context effects on answers to attitude questions. 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. 111–131). New York, NY: John Wiley & Sons.

Tourangeau, R. (2000). Remembering what happened: Memory errors and survey reports. In A. A. Stone, C. A. Bachrach, J. B. Jobe, H. S. Kurtzman, & V. S. Cain (Eds.), *The science of self-report: Implications for research and practice* (pp. 29–48). Mahwah, NJ: Lawrence Erlbaum.

Tourangeau, R. (2018). The survey response process from a cognitive viewpoint. *Quality Assurance in Education*, *26*(2), 169–181. Retrieved from https://www.emerald.com/insight/content/doi/10.1108/QAE-06-2017-0034/full/html

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

Tourangeau, R., Couper, M. P., & Conrad, F. G. (2004). Spacing, position, and order: Interpretive heuristics for visual features of survey questions. *Public Opinion Quarterly*, *68*(3), 368–393. https://doi.org/10.1093/poq/nfh035

Tourangeau, R., Maitland, A., Steiger, D., & Yan, T. (2020). A framework for making decisions about question evaluation methods. 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. 47–74). Hoboken, NJ: John Wiley & Sons.

Tourangeau, R., & Rasinski, K. A. (1988). Cognitive processes underlying context effects in attitude measurement. *Psychological Bulletin*, *103*, 299–314. https://doi.org/10.1037/0033-2909.103.3.299

Tourangeau, R., Rasinski, K. A., & Bradburn, N. M. (1991). Measuring happiness in surveys: A test of the subtraction hypothesis. *Public Opinion Quarterly*, *55*(2), 255–266. https://doi.org/10.1086/269256

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

Tourangeau, R., Singer, E., & Presser, S. (2003). Context effects in attitude surveys: Effects on remote items and impact on predictive validity. *Sociological Methods & Research*, *31*(4), 486–513. https://doi.org/10.1177/0049124103251950

Tourangeau, R., & Yan, T. (2007). Sensitive questions in surveys. *Psychological Bulletin*, *133*(5), 859–883. https://doi.org/10.1037/0033-2909.133.5.859

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

Van Vaerenbergh, Y., & Thomas, T. D. (2013). Response styles in survey research: A literature review of antecedents, consequences, and remedies. *International Journal of Public Opinion Research*, *25*(2), 195–217. https://doi.org/10.1093/ijpor/eds021

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. (1994). Cognitive interviewing and questionnaire design: A training manual. *Cognitive Methods Staff Working Paper Series 7*, 1–56.

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. (2015b). Research synthesis: The practice of cross-cultural cognitive interviewing. *Public Opinion Quarterly*, *79*(S1), 359–395. https://doi.org/10.1093/poq/nfu092

Willis, G. B. (2020). Questionnaire design, development, evaluation, and testing: Where are we, and where are we headed? 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. 1–23). Hoboken, NJ: John Wiley & Sons.

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

Willis, G. B., DeMaio, T., & Harris-Kojetin, B. (1999). Is the bandwagon headed to the methodological promised land? Evaluating the validity of cognitive interviewing techniques. 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. 133–153). New York, NY: John Wiley & Sons.

Willis, G. B., & Schechter, S. (1997). Evaluation of cognitive interviewing techniques: Do the results generalize to the field? *Bulletin de Méthodologie Sociologique*, *55*, 40–66. https://doi.org/10.1177/075910639705500105

Willis, G. B., Schechter, S., & Whitaker, K. (1999). A comparison of cognitive interviewing, expert review, and behavior coding: What do they tell us? Retrieved from http://www.asasrms.org/Proceedings/papers/1999_006.pdf

Willson, S., Cibelli Hibben, K., & Gregory-Lee, K. (2022). *Results from a cognitive interview evaluation of a subset of questions for the National Intimate Partner and Sexual Violence Survey: Round 2*. Hyattsville, MD. Retrieved from https://wwwn.cdc.gov/qbank/report/Willson_2022_NCHS_NISVS_Round2.pdf

Willson, S., & Miller, K. (2022). *Cognitive interview evaluation of demographic questions for the US Department of State Global Employee Management System*. Hyattsville, MD. Retrieved from https://wwwn.cdc.gov/qbank/report/Willson_2023_NCHS_GEMS.pdf

Willson, S., Scanlon, P. J., & Miller, K. (2022). Question evaluation for real-time surveys: Lessons from COVID-19 data collection. *SSM. Qualitative Research in Health*, *2*, Online first. https://doi.org/10.1016/j.ssmqr.2022.100164

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.

World Values Survey (2014). *World values survey. Wave six*. Retrieved from http://www.worldvaluessurvey.org/WVSDocumentationWV6.jsp

Wundt, W. (1907). Über Ausfrageexperimente und über Methoden zur Psychologie des Denkens. *Psychologische Studien*, *3*, 301–360.

Wundt, W. (1908). Kritische Nachlese zur Ausfragemethode. *Archiv für die gesamte Psychologie*, *11*, 445–459.

Yan, T., & Curtin, R. (2010). The relation between unit nonresponse and item nonresponse: A response continuum perspective. *International Journal of Public Opinion Research*, *22*(4), 535–551. https://doi.org/10.1093/ijpor/edq037

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, NJ: John Wiley & Sons.

Yan, T., Kreuter, F., & Tourangeau, R. (2012). Evaluating survey questions: A comparison of methods. *Journal of Official Statistics*, *28*(4), 503–529. Retrieved from https://www.scb.se/contentassets/ff271eeeca694f47ae99b942de61df83/evaluating-survey-questions-a-comparison-of-methods.pdf

Yan, T., & Olson, K. (2013). Analyzing paradata to investigate measurement error. In F. Kreuter (Ed.), *Improving surveys with paradata: Analytic uses of process information* (pp. 73–95). Hoboken, New Jersey: John Wiley & Sons.

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., Williams, D., Maitland, A., & Tourangeau, R. (2016). *Use of eye-tracking to measure response burden.* AAPOR, Austin, Texas.

Yu, E., Fobia, A. C., Graber, J., Holzberg, J., Kaplan, R., Kopp, B., . . . Scanlon, P. J. (2019). *White Paper: Experiences using online testing to support survey-methods research and pre-testing in the federal government*. U.S. Census Bureau: Research and Methodology Directorate, Center for Behavioral Science Methods Research Report Series (Survey Methodology #2019-06). Retrieved from https://www.bls.gov/osmr/research-papers/2019/pdf/st190010.pdf

Zanna, M. P., Olson, J. M., & Fazio, R. H. (1981). Self-perception and attitude-behavior consistency. *Personality and Social Psychology Bulletin*, *7*(2), 252–256. https://doi.org/10.1177/014616728172011

Zuell, C. (2016). *Open-ended Questions*. *GESIS Survey Guidelines*. Mannheim, Germany: 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

Zuell, C., & Scholz, E. (2015). Who is willing to answer open-ended questions on the meaning of left and right? *Bulletin de Méthodologie Sociologique*, *127*(1), 26–42. https://doi.org/10.1177/0759106315582199