*Journal of Survey Statistics and Methodology (2022) **10**, 923–944*

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

KATHARINA MEITINGER*
ADRIAN TOROSLU
KLARA RAIBER
MICHAEL BRAUN

Web probing is a valuable tool to assess the validity and comparability of survey items. It uses different probe types—such as category-selection probes and specific probes—to inquire about different aspects of an item. Previous web probing studies often asked one probe type per item, but research situations exist where it might be preferable to test potentially problematic items with multiple probes. However, the response behavior might be affected by two factors: question order and the visual presentation of probes on one screen versus multiple screens as well as their interaction. In this study, we report evidence from a web experiment that was conducted with 532 respondents from Germany in September 2013. Experimental groups varied by screen number (1 versus 2) and probe order (category-selection probe first versus specific probe first). We assessed the impact of these manipulations on several indicators of response quality, probe answer content, and the respondents' motivation with logistic regressions and two-way ANOVAs. We reveal that multiple mechanisms push response behavior in this context: perceived response burden, the focus of attention, the need for justification, and verbal context effects. We find that response behavior in the condition with two screens and category-selection probe first outperforms all other

KATHARINA MEITINGER is an Assistant Professor and ADRIAN TOROSLU is a PhD candidate at Utrecht University. KLARA RAIBER is a PhD candidate at the Radboud University in Nijmegen. MICHAEL BRAUN is Senior Project Consultant at GESIS – Leibniz-Institute for the Social Sciences and Adjunct Professor at the University of Mannheim. This work was supported by the German Research Foundation (DFG) as part of the project "Optimizing Probing Procedures for Cross-National Web Surveys" [BR 908/5-1 to Michael Braun, Wolfgang Bandilla, and Lars Kaczmirek].

*Address correspondence to Katharina Meitinger, Utrecht University; E-mail: k.m.meitinger@uu.nl.

doi: 10.1093/jssam/smaa043 Advance access publication 12 June 2021

experimental conditions. We recommend this implementation in all but one scenario: if the goal is to test an item that includes a key term with a potentially too large lexical scope, we recommend starting with a specific probe but on the same screen as the category-selection probe.

KEYWORDS: Multiple probes; Question order; Visual design; Web Probing.

1. INTRODUCTION AND RESEARCH QUESTIONS

Web probing is a crucial tool to assess the validity and comparability of survey questions (Behr, Meitinger, Braun, and Kaczmirek 2017, 2019) and research applying this method is increasing due to the valuable methodological and substantive insights this approach can provide (e.g., Behr, Bandilla, Kaczmirek, and Braun 2014; Meitinger 2017, 2018; Braun, Behr, and Díez Medrano 2018; Efremova, Panyusheva, Schmidt, and Zercher 2018; Schulz, Meitinger, Braun, and Behr 2018; Braun, Behr, Meitinger, Raiber, and Repke 2019; Lee, McClain, Behr, and Meitinger 2020). The method of web probing applies probing techniques from cognitive interviewing in web surveys. Probes are questions that ask respondents to provide additional information after having answered a closed item (Beatty and Willis 2007). In previous research, respondents typically received one probe on a separate screen directly after responding to the item that needed to be tested (Braun, Behr, Kaczmirek, and Bandilla 2015). Figure 1 shows the typical implementation of web probing.

Different probe types can address different aspects of the answer process. For example, (1) a *category-selection probe* asks respondents for the reasons why a certain answer category has been chosen; (2) a *specific probe* requests respondents to provide additional information on a particular detail of an item; and (3) a *comprehension probe* requests a definition of a specific term (Prüfer and Rexroth 2005; Willis 2005).

1.1 Previous Research

Previous web probing studies predominantly evaluated each closed item with *one* probe question (Meitinger, Braun, and Behr 2018). In the web mode, researchers need to decide on the different probes in advance and have to program them before data collection because web probing is not as interactive as traditional cognitive interviewing (Meitinger and Behr 2016) where the interviewer can ask additional spontaneous and emergent probes at any time (Willis 2005) that are adapted to the interview situation. Therefore, multiple probes might be preferable in situations where questionnaire designers are uncertain which aspect of an item might be problematic or where different issues might appear at different stages of the question–answer process (e.g., issues related to comprehension or retrieval). For example, in a question asking respondents

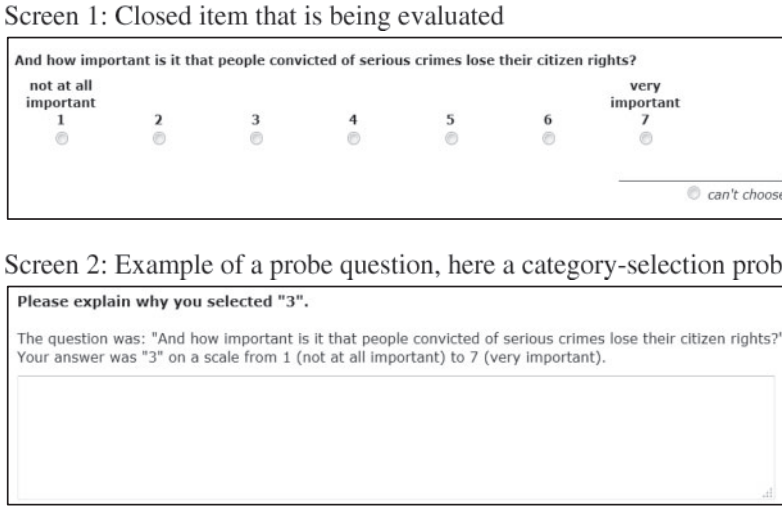

Figure 1. Illustration of Web Probing Procedure.

how important it is for them that people convicted of serious crimes lose their citizen rights, several aspects could be challenging for the respondent: some respondents might differ in the definition of what constitutes a "serious crime" (Schulz et al. 2018), some might struggle to understand the vague term "citizen rights," and some might apply different reasoning for choosing a particular answer category. Therefore, a comprehension probe, a specific probe, and a category-selection probe could provide valuable insights into the various aspects of this question (Meitinger et al. 2018).

Despite the potential benefits of asking multiple probes, there is only scarce empirical evidence concerning their optimal implementation and impact on different aspects of response behavior: response quality, answer content, and respondents' motivation. Meitinger et al. (2018) analyzed the impact of probe sequence (category-selection probe followed by a specific and a comprehension probe versus comprehension probe followed by a specific and a category-selection probe) on response behavior in five countries. They found that the probe sequence did not affect response length. However, probe nonresponse increased for subsequent probes when respondents first received a comprehension probe, but this was the case only in Great Britain and the United States. The percentage of mismatching answers (answers that do not contain the expected information) at the comprehension probe was higher in all countries when this was the first probe. Respondents' motivation was negatively affected when the category-selection probe appeared as the third probe. Answer content for the category-selection and the specific probes were unaffected by the sequence, and for the comprehension probe, they found a significant difference for Germany and Mexico only. Obviously, the different indicators used did not

Figure 2. Design Options for Multiple Probes: Multiple- Versus Single-Screen Design.

uniformly point in the same direction. The authors argued that this might be due to divergent indicator quality; that is, whether long answers and many themes are per se an advantage is debatable (see also Meitinger, Behr, and Braun 2019).

Meitinger and colleagues did not manipulate the question order of specific probes and also did not manipulate the number of screens on which the multiple probes appear. Therefore, the visual presentation of multiple probes on one screen versus multiple screens and its interaction with probe order has not been studied, yet. Figure 2 shows the alternative to implement multiple probes on either multiple screens or one screen.

1.2 Research Questions

In this article, we aim at answering the following research questions:

(1) Does the visual presentation of probes on one single versus multiple separate screens affect response behavior?
(2) Does the order in which different probe types appear have an impact on response behavior?
(3) Is there an interaction effect between probe order and the number of screens on response behavior?

1.3 Effects of Verbal and Visual Information

When considering the optimal implementation of multiple probes, it is necessary to distinguish between effects induced by visual information and by verbal information (Ware 2000, see also Redline, Dillman, Dajani, and Scaggs 2002; Couper, Tourangeau, and Kenyon 2004). The *visual information* of probes

consists of the multiple answer boxes that could either be placed on one screen or multiple screens. The *verbal information* of a probe is the question wording (the actual probe), the repetition of the closed item, and—for category-selection probes—the repetition of the selected answer category. Due to different mechanisms, manipulations of the visual and verbal information have consequences on response quality, the respondents' motivation, and the answer content.

1.3.1 The impact of increased perceived response burden on response quality and answer content. Open-ended questions, such as probes, impose a higher response burden on respondents than closed questions (Bradburn 1978) because they have to formulate their answers in their own words (Keusch 2014). The respondents cannot rely on pre-defined response categories to infer the question meaning (Dillman, Smyth, and Christian 2009) or find themes they may not have considered otherwise (Schwarz 1999). As a consequence, open-ended questions are potentially more affected by issues of response quality (e.g., higher item nonresponse) than closed items. In web surveys, a motivating interviewer is missing for such 'burdensome' questions (Meitinger and Behr 2016). By asking *multiple* probes, the imposed response burden further increases, and the *perceived* response burden is particularly high if multiple open-ended questions appear on the same screen (Smyth, Dillman, and Christian 2007). This might tempt respondents to refuse to start the cognitive response process of question comprehension, retrieval of relevant information, forming of judgment, and reporting of the response (Tourangeau, Rips and Rasinski 2000, see also Meitinger and Kunz, 2018). Alternatively, it might trigger respondents to switch to satisficing behavior by executing the cognitive stages less thoroughly (weak satisficing) or completely skip one or more stages (strong satisficing) (Krosnick 1991). The perceived response burden, thus, has consequences for response quality. Considering these aspects, we derive the following hypotheses:

> H1: Due to increased perceived response burden, response quality decreases (nonresponse increases) at the second probe if respondents receive multiple probes per screen than when they receive one probe per screen (Main effect number of screens).

Besides a reduced response quality, respondents might also be less willing to write all aspects that they have thought of for both probes if they receive multiple probes on the same screen. The response task might just seem too burdensome which reduces the answer content of the second probe.

> H2: Due to increased perceived response burden, answer content decreases at the second probe if respondents receive multiple probes per screen than when they receive one probe on multiple screens (Main effect number of screens).

1.3.2 The relation of focus of attention, the need to justify, and mismatches and their impact on response behavior. A further relevant factor concerning visual information might be that respondents tend to focus their attention on a very narrow region on the screen (Kahneman 1973). To ensure that respondents see relevant information without having to move their eyes, it should appear within this region (Christian, Dillman, and Smyth 2007). The vision of the respondents on the screen might not even cover one probe (Kahneman 1973). Placing two probes on one screen might increase the risk that a part of the relevant information is outside the respondents' view even more. As a consequence, the respondents might not pay attention to the question text of the second probe. Alternatively, respondents could satisfice in this context. If respondents receive two probes on the same screen, they have to divide their attention and might be tempted to read the second probe less attentively. Previous research regarding grid design in web surveys indicated that a reduction in cognitive load potentially reduces satisficing by increasing attention to the task (Couper et al. 2013). The cognitive load reduces if respondents receive only one probe per screen.

One consequence of a differential focus of attention is that respondents receiving both probes on one screen will read the second probe less thoroughly which could impact the answer content of both probes by providing mismatching responses. A mismatch occurs, for example, when a respondent answers a category-selection probe (e.g., explains the reasons for answer selection) at a specific probe (Behr et al. 2014; Meitinger and Behr 2016; Meitinger et al. 2018).

Verbal information can also affect mismatching behavior. Behr et al. (2014) found that respondents often explain their motivation for their answer selection at a closed item (i.e. to respond to a category-selection probe) rather than elaborate on the things that came to their minds when reading an item (i.e. to respond to a specific probe). This means that respondents often feel the urge to justify their answers even if they are not asked for it. We call this effect the need for justification. Thus, respondents might show a tendency to report their justification for a response selection even if it is not requested, that is, in the case of a specific probe (Meitinger et al. 2018). This tendency should be most pronounced if both probes appear on separate screens. In this case, respondents must assume that they will not get a second chance to communicate the reason for their answer choice at the closed item if they receive a specific probe as the first probe.

> H3: Due to reduced focus of attention and the need for justification, respondents receiving probes on multiple screens and as a first probe a specific probe provide more mismatches which reduces the answer content compared to respondents that receive a category-selection probe as first probe or their probes on the same screen (Interaction effect).

Mismatches at the first probe also have consequences for response behavior at the second probe. Respondents who already provided a mismatching response at the first probe, when they answered a specific probe as if it were a category-selection probe, might not be willing to respond to the second probe when they receive the actual category-selection probe. These respondents are more likely to show an elevated level of nonresponse. Mismatch at the first probe, however, is more likely when both probes are on different screens because, in this case, respondents cannot see that the category-selection probe (which they "prefer") is still coming.

H4: Due to mismatches, respondents receiving probes on multiple screens and as a first probe a specific probe provide more responses with reduced response quality at the second probe than respondents that receive a category-selection probe as first probe or their probes on the same screen (Interaction effect).

Mismatch responding at the first probe might also frustrate respondents and reduce their motivation (e.g., respondents that explained the reasons for their answer selection at a specific probe might be irritated to have to repeat these reasons at the actual category-selection probe).

H5: Respondents receiving probes on multiple screens and as a first probe a specific probe provide more mismatches at the first probe which reduces their motivation at the second probe compared to respondents that receive a category-selection probe as first probe or their probes on the same screen (Interaction effect).

1.3.3 The impact of question order effects on answer content. When asking multiple probes, a question order effect can appear that could have a clarifying effect on respondents. Question order effect means that previous questions affect how respondents interpret and answer subsequent questions. These effects are more likely to occur if questions are very close to each other in topic and location in the survey (Smyth, Dillman, and Christian 2008) and if the previous question triggers associations or thoughts that are easier accessible for later interpretation and response to the following question (Schwarz and Bless 1992; Smyth et al. 2008). Respondents might rely for their interpretation of the second probe on the verbal information of the first probe. For example, respondents might use the information and thought processes from answering a specific probe when responding to a category-selection probe. Respondents might do this if they perceive the second probe as ambiguous since respondents tend to draw on the context to determine the meaning of ambiguous questions (Schwarz and Strack 1991). Vice versa, respondents might find a response more difficult when they receive the more ambiguous probe as the first probe. A category-selection probe is potentially more ambiguous than a specific probe if the closed items contain vague key terms. Specific probes

clarify the expected response content (e.g., group of immigrants). In contrast, category-selection probes ask for the reasons for selecting a response category. If respondents already struggled with the ambiguity of the closed item, they will find it difficult to define the scope of the category-selection probe. However, if they receive a specific probe as the first probe, they can define some of the key terms (e.g., they thought of Italians and Turks) and apply this scope to their response to the category-selection probe.

H6: Due to question order effects, respondents experience less uncertainty about the expected answer scope at the category-selection probe if they receive it after a specific probe than respondents that receive the category-selection probe as first probe (Main effect probe order).

2. DATA AND METHODS

2.1 Sample

The experimental study was embedded in a web survey that was conducted in September 2013 among participants from a nonprobability online access panel in Germany. The panel provider was respondi (https://www.respondi.com/EN/), a company that adheres to ISO 26362, an international standard to raise quality and transparency in access panels in market, opinion, and social research. The sample was based on quotas for age (18–30, 31–50, and 61–65), gender, and education (lower and higher). All quotas were met (see Appendix A.1). The sample was limited to desktop users only. From the 1,005 panelists invited to the web survey, 404 were screened out because respective quotas were already full. In total, 532 respondents completed the survey with a break-off rate (Callegaro and DiSogra 2008) of 11.48 percent. Among all respondents, 50 percent were female and the average age was 42 years. On average, the questionnaire took 18.61 minutes to complete (Mdn = 14.49).[1]

2.2 Experimental Design

We selected an item from the battery on general national pride from the 2013 National Identity module of the International Social Survey Program (ISSP Research Group 2015). The item was "The world would be a better place if people from other countries were more like the Germans." Response alternatives were "strongly agree," "agree," "neither agree nor disagree," "disagree," and "strongly disagree." A "can't choose" category was offered.

In our web survey, this item battery was part of a longer questionnaire replicating questions from the ISSP modules on National Identity as well as Family

1. The data set is available from the authors upon request.

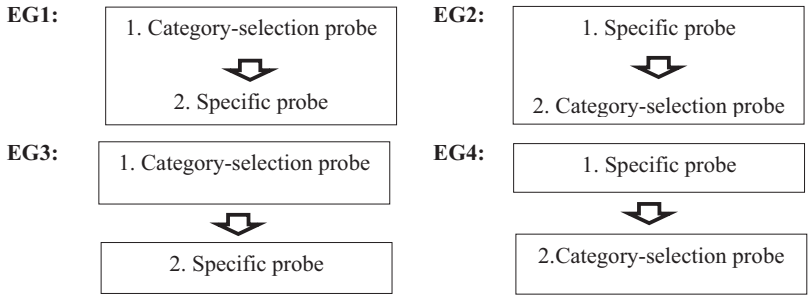

Figure 3. Experimental Design.

and Gender Roles. The experiment was implemented near the beginning of the questionnaire. Based on team discussions, we identified two potential issues for this item: Respondents could apply very different reasoning for selecting a response category and respondents might have very different countries in mind when they respond to this question. Therefore, we decided to ask a category-selection probe (reasons for selecting an answer category) and a specific probe ("Which countries were you thinking about when you were answering the question?") to assess this item.

We manipulated two aspects in our experimental setting: the probe order and the number of screens. Respondents were randomly assigned to four experimental conditions. Experimental group 1 (EG1: 1 screen—CSP 1st) received on one screen first a category-selection probe and then the specific probe. Experimental group 2 (EG2: 1 screen—SP 1st) received first the specific probe and then the category-selection probe on a single screen. Experimental group 3 (EG3: 2 screens—CSP 1st) received the probes in the same order as group 1 (i.e., the category-selection probe followed by the specific probe) but on two different screens. Finally, experimental group 4 (EG4: 2 screens—SP 1st) received the probes in the same order as EG 2 (i.e., the specific probe followed by the category-selection probe) but on two different screens (see figure 3).

2.3 Coding Procedure

Based on the responses to the specific and category-selection probes, we developed two separate coding schemata that captured the different themes mentioned but also methodological aspects (e.g., mismatching response, reduced motivation, nonresponse).[2] Two student assistants received training for each coding schema and coded all responses to the probes. Inter-coder reliability calculated according to Holsti (1969) was deemed satisfactory (specific: 88

2. The full coding schemata are available from the authors upon request.

Table 1. Relation of Hypotheses, Effect Types, Aspects of Response Behavior, and Indicators

Hypothesis	Effect	Aspect of response behavior	Indicator
H1	Main effect number of screens	Response quality	Nonresponse
H2	Main effect number of screens	Answer content	Number of themes
H3	Interaction effect	Answer content	Mismatches
H4	Interaction effect	Response quality	Nonresponse
H5	Interaction effect	Respondents' motivation	Overt signs of reduced motivation
H6	Main effect order of probes	Answer content	"It depends" responses

percent; category selection: 82 percent). Any coding discrepancies were corrected in the final dataset.

2.4 Measures

To assess the impact of the number of screens and probe order on response behavior, we measured response quality with probe nonresponse. We assessed answer content with three indicators: number of themes mentioned, mismatches, and "it depends" responses. Respondents' motivation was assessed with the indicator overt signs of reduced motivation. (1) For *probe nonresponse*, we coded all respondents as nonrespondents that either provided an empty answer box or wrote a non-substantive response, such as unintelligible letter combinations (e.g., "xcvbnm"), explicit refusals (e.g., "n/a," "no comment"), don't knows, and meaningless or incomprehensible answers. (2) We measured the *number of themes* as the number of substantive themes mentioned. (3) The indicator *mismatching probe response* flags respondents who wrote answers to a different probe type than required. For example, respondents mentioned reasons for choosing a certain answer value at the probe where they were supposed to report their associations for a key term, that is, they treated a specific probe as if it were a category-selection probe. (4) The indicator *"it depends" responses* captures respondents that were uncertain about the scope of key terms at the category-selection probe and, as a consequence, struggled with their probe response. For example, respondents pointed out that their response would differ depending on which countries they would use as a comparison. (5) The indicator *overt signs of reduced motivation* flags respondents who complained that they had already answered the question (e.g., "don't ask the same question twice."). Table 1 summarizes which indicator measures which aspect of response behavior and which hypothesis and effect are tested.

2.5 Analysis

All analyses were conducted using Stata version 14. We used the full sample for the analysis of probe nonresponse. All remaining analyses were limited to respondents who provided a substantive answer to the probes. We report descriptive statistics for each indicator overall and by experimental group. To assess differences between the effects of probe order, the number of screens, and their interaction, a two-factorial analysis of variance (ANOVA) was used for continuous dependent variables. For categorical indicators, we report Pearson's chi-square or Fisher's exact tests and the results of logistic regressions with the predictors of probe order, number of screens, and their interaction if the interaction term was significant.

3. RESULTS

3.1 Response Quality: Nonresponse

For the indicator nonresponse, we report nonresponse by probe type as well as the overall nonresponse (probes combined) (see table 2). In addition, we have to consider that respondents will be, in any case, more prone to nonresponse at the second probe. It is at this point that response burden is most likely to become manifest in its consequences.

To address our hypothesis that perceived response burden is particularly high if respondents receive both probes on one screen (H1), we need to assess whether respondents that received a specific probe as second probe provided more nonresponse when they were in the one-screen condition (EG1) than in the two-screen condition (EG3). The same applies to the category-selection probe (EG2 versus EG4). Indeed, respondents that received a specific probe as a second probe provided more nonresponse when they received all probes on one screen (EG1, 18.05 percent) than respondents that received the specific probe on a separate screen (EG3, 10.22 percent). However, this difference is not significant [χ^2 (1, $N = 270$) = 3.42, $p = .064$]. Respondents that received a category-selection probe as a second probe provided less nonresponse in the one-screen condition (EG2, 11.45 percent) than in the two-screen condition (EG4, 19.08 percent). While this is the opposite of what we expected, the difference is not significant [χ^2 (1, $N = 262$) = 2.95, $p = .086$]. Therefore, we cannot confirm H1.

Instead, a more complex nonresponse pattern emerged for both, category-selection and specific probe. Nonresponse increased if a category-selection probe was asked first on one screen (EG1) or if a specific probe was asked first in the two-screen condition (EG4). In both settings, the nonresponse was elevated for the first but also the second probe. In contrast, nonresponse was comparatively lower at the second probe if respondents had already responded to a specific probe and then received a category-selection probe on the next screen

Table 2. Percentage of Nonresponse by Experimental Group and Probe Type and Overall

				CSP		SP		Overall NR	
EG	Screen	Order	*N*	*n*	%	*n*	%	*n*[a]	%
Overall			532	76	14.29	74	13.91	98	18.42
1	1 screen	CSP 1st	133	22	16.54	24	18.05	31	23.31
2		SP 1st	131	15	11.45	12	9.16	18	13.74
3	2 screens	CSP 1st	137	14	10.22	14	10.22	19	13.87
4		SP 1st	131	25	19.08	24	18.32	30	22.90
				χ^2 (3, $N = 532$) = 5.73, $p = .126$, Cramer's $V = 0.10$		χ^2 (3, $N = 532$) = 8.05, $p = .045$, Cramer's $V = 0.12$		χ^2 (3, $N = 532$) = 7.66, $p = .054$, Cramer's $V = 0.12$	

NOTE.—EG, experimental group; CSP, category-selection probe; SP, specific probe NR, nonresponse.
[a]*n* refers to number of respondents who gave at least one nonresponse at either the category-selection probe or the specific probe.

Table 3. Logit Regression with Odds Ratios for Multiple Indicators with Number of Screens, Order (and Interaction) as Predictors

	(1) Overall nonresponse	(2) Mismatch at specific probe	(3) Motivation	(4) "It depends"
Screen (reference: 1 screen)				
2 screens	0.53*	1.51**	0.57	1.01
	(0.17)	(0.10)	(0.33)	(0.21)
Order (reference: CSP 1st)				
SP 1st	0.52*	0.50	3.79*	0.64*
	(0.17)	(0.22)	(2.50)	(0.13)
Screen × order				
2 screens × SP 1st	3.52**	20.05***		
	(1.62)	(15.54)		
Constant	0.30***	0.17***	0.02***	0.58**
	(0.06)	(0.05)	(0.01)	(0.10)
N	532	434	434	434
Adj. R^2	0.02	0.08	0.05	0.01

NOTE.— Standard errors are in parentheses.
*$p < .05$, **$p < .01$, ***$p < .001$. Interactions listed only if they were significant. CSP, category-selection probe; SP, specific probe.

(EG2). In a similar vein, nonresponse was reduced if respondents started with a category-selection probe in the two-screen condition (EG3).

Overall, we find a powerful interaction effect between the number of screens and the probe order when assessing the overall nonresponse (H4 confirmed; see Table 3, Model 1). Overall nonresponse was lowest if either the specific probe was asked first on one screen (EG2: 13.74 percent) or the category-selection probe was asked first on two screens (EG3: 13.87 percent). In contrast, it was highest when either the category-selection probe was asked first on one screen (EG1: 23.31 percent) or the specific probe was asked first on two screens (EG4: 22.90 percent).

3.2 Answer Content: Number of Themes

Table 4 presents the average number of themes mentioned in substantive responses by probe type and experimental condition. On average, respondents gave fewer reasons for answer selection at the category-selection probe than themes at the specific probe. Due to the different tasks involved in both probe types, this is not surprising. It is easier to think about—or retrieve—several countries which the respondents had in mind than reporting multiple reasons for an answer selection. When considering the combined average number of

Table 4. Number of Mentioned Themes by Experimental Group and Probe Type

EG	Screen	Order	*N*	CSP	SP	Combined
				Mean (SD)	Mean (SD)	Mean (SD)
1	1 screen	CSP 1st	102	1.27 (0.82)	1.41 (1.28)	2.69 (1.75)
2		SP 1st	113	1.15 (0.64)	1.51 (1.10)	2.66 (1.41)
3	2 screens	CSP 1st	118	1.25 (0.68)	1.78 (1.39)	3.03 (1.70)
4		SP 1st	101	1.35 (0.79)	1.41 (1.47)	2.75 (1.98)
Overall			434	1.25 (0.73)	1.54 (1.32)	2.79 (1.71)

NOTE.— EG, Experimental group; CSP, category-selection probe; SP, specific probe.

themes, respondents receiving both probes on one screen (EG1 and EG2) mentioned fewer themes than respondents receiving probes on two screens (EG3 and EG4). Respondents in EG3 provided the most and respondents in EG2 provided the fewest themes.

For the indicator number of themes, we were particularly interested in whether an increased perceived response burden in a multiple-probe per screen setting pushes respondents to report fewer themes at the second probe (H2). To address our hypothesis, we needed to assess whether respondents that received a specific probe as the second probe provided fewer themes when they were in the one-screen condition (EG1) compared to the two-screen condition (EG3). The same applies to the category-selection probe (EG2 versus EG4).

Indeed, respondents that received a specific probe as a second probe mentioned fewer themes when they received both probes on one screen (EG1) compared to respondents that received the specific probe on a separate screen (EG3). This difference is significant with a small effect size [one-tailed $t(218) = -2.03$, $p = .022$, $r = 0.14$]. In a similar vein, respondents that received a category-selection probe as a second probe mentioned fewer themes in the one-screen condition (EG2) than in the two-screen condition (EG4). Once again, this difference is significant with a small effect size [$t(212) = -1.99$, $p = .024$, $r = 0.14$]. Therefore, we can confirm our H2.

3.3 Answer Content: Mismatching Probes

Concerning the indicator mismatching probes, we expected an interaction effect due to the differential focus of attention and the need for justification (H3). Table 5 summarizes the occurrence of mismatching probe responses by experimental groups. When a category-selection probe was asked, none of the respondents answered as if a specific probe had been presented. In contrast, mismatches appeared when respondents had to answer a specific probe but instead provided a response as if a category-selection probe had been presented.

Table 5. Percentage of Mismatches at Specific Probe by Experimental Group

EG	Screen	Order	*N*	*n*	%
Overall			532	51	9.59
1	1 screen	CSP 1st	133	16	12.03
2		SP 1st	131	9	6.87
3	2 screens	CSP 1st	137	4	2.92
4		SP 1st	131	22	16.79
				χ^2 (3, $N = 532$) = 16.91, $p = .001$, Cramer's $V = 0.18$	

NOTE.— EG, experimental group; CSP, category-selection probe; SP, specific probe.

Therefore, the table only considers mismatching at the specific probe. In EG1 (1-screen—CSP 1st), 12.03 percent of the respondents did not realize that the second probe was a specific probe and provided a second reason for selecting an answer category or provided some justification for their answer to the category-selection probe. In contrast, mismatches were lower if respondents first received the specific probe in the one-screen condition (EG2: 1-screen—SP 1st: 6.87 percent). For the two-screen setting, mismatches were lowest if respondents received the specific probe on the second screen (EG3: 2.92 percent) and highest if they received it on the first screen (EG4: 16.79 percent). In the logistic regression, we found a highly significant interaction effect between the number of screens and the probe order (OR: 20.05) (see Table 3, Model 2). Therefore, we can confirm our Hypothesis 4. Only two out of the fifty-one respondents that gave a mismatch at the specific probe already mentioned which country they had in mind when responding at the category-selection probe. As a consequence, mismatches reduce the completeness of the overall response, in the sense that the response to the specific probe is completely missing.

3.4 Respondents' Motivation: Overt Signs of Reduced Motivation

Mismatching behavior can also impact the motivation of respondents (H5; see Table 6). Respondents' overt signs of a reduced motivation at the category-selection probe are most frequent when they received it as the second probe (EG2 and EG4). In contrast, respondents that had received the category-selection probe first did show either less (EG1: 1 screen—CSP 1st) or no (EG3: 2 screens—CSP 1st) indications of reduced motivation at the specific probe. The overall difference between the experimental groups is statistically significant (*two-sided Fisher's exact test* = 0.039). In the logistic regression, the predictor of probe order was significant, indicating that receiving first a specific probe increased the odds that a respondent complained at the category-

Table 6. Percentage of Signs of Reduced Motivation by Experimental Condition (Both Probe Types Combined)

EG	Screen	Order	*N*	*n*	%
Overall			434	14	9.59
1	1 screen	CSP 1st	102	3	2.94
2		SP 1st	113	6	5.30
3	2 screens	CSP 1st	118	0	0
4		SP 1st	101	5	4.95
				Fisher's exact = 0.039	

NOTE.— EG, experimental group; CSP, category-selection probe; SP, specific probe.

selection probe (see Table 3, Model 3). Neither the number of screens nor the interaction of screen and order was significant (H5 not confirmed). It is important to note that the overall prevalence of this indicator is very low. However, these findings are in line with the results of Meitinger et al. (2018).

Only one out of the fourteen respondents that showed signs of reduced motivation provided the response to the specific and category-selection probe at the first probe. Therefore, overt signs of reduced motivation are predominantly a consequence of a mismatching response at the first probe. As a consequence, signs of reduced motivations reduce the completeness of responses because respondents only provide the information for the category-selection probe but not for the specific probe.

3.5 Answer Content: "It Depends" Responses

Besides the number of themes and mismatches, another aspect regarding answer content is relevant here: Does the content of the first probe affect the interpretation of the second probe? That is, are context effects present (H6)? The item of this study had key words with potentially large lexical scopes (item wording: "The world would be a better place if people from other countries were more like the Germans"). For example, the reason for selecting a response category could differ depending on which country and which group of people (e.g., specific immigrant groups) the respondent had in mind. Several respondents struggled with this large lexical scope and pointed out that opinions would differ depending on the point of comparison. When comparing these "it depends" responses for the category-selection probe by experimental group, we see a striking pattern (see Table 7): more respondents provided "it depends" responses if they received the category-selection probe as the first probe. Although the overall effect is not significant [χ^2 (3, $N = 532$) = 4.75, p = .191], the predictor for probe order is significant in the logistic regression (see Table 3, Model 4, H6 confirmed).

Table 7. Percentage of Respondents Mentioning "It Depends" at CSP

EG	Screen	Order	*N*	*n*	%
Overall			434	139	32.03
1	1 screen	CSP 1st	102	38	37.25
2		SP 1st	113	30	26.55
3	2 screens	CSP 1st	118	43	36.44
4		SP 1st	101	28	27.72
				χ^2 (3, $N = 532$) = 4.75, $p = .191$, Cramer's $V = 0.10$	

NOTE.— EG, experimental group; CSP, category-selection probe; SP, specific probe.

4. DISCUSSION AND CONCLUSION

We analyzed the impact of probe sequence (category-selection probe at the first versus the second position) and the presentation of probes (category-selection and specific probes) on one or two separate screens on response behavior. Extending research of Meitinger et al. (2018), we distinguished between the impact of visual (number of screens) and verbal information (probe order) on response behavior and also assessed their interaction. We assumed that different mechanisms, such as the perceived response burden, the focus of attention, the need for justification, and context effects, could impact response behavior. We distinguished between three aspects of response behavior: response quality, answer content, and respondents' motivation. We used the indicator of nonresponse to measure the impact on response quality. For answer content, we drew on the indicators number of themes, mismatches, and "it depends" responses. Finally, the indicator overt signs of reduced motivation served as an indicator of respondents' motivation.

Based on the identified mechanisms, we developed six hypotheses of how the number of screens and probe order could impact the different aspects of response behavior. From a visual point of view, we were concerned that respondents perceive the response burden as higher when they received two probes on the same screen than one probe on two screens. This might lead to a decrease in response quality at the second probe. Concerning response quality, perceived response burden does not seem to be the only mechanism at work. For nonresponse, we found an increase for the specific probe if respondents received this probe in the one-screen condition. However, we found the opposite effect for the category-selection probe. Nonresponse was lower in the one-screen condition than in the two-screen condition if this probe was asked in second position. Both differences were not significant. Nonresponse seems to be driven by a complex interaction of number of screens and probe order which indicates that perceived response burden is not the only driving force of nonresponse in this context.

Perceived response burden can also impact answer content, namely the number of themes mentioned at the second probe. For both probe types, we found that respondents mentioned significantly fewer themes at the second probe if they received two probes on one screen (EG1 and EG2). Respondents seem to perceive this setting as burdensome and reduce the response effort by writing fewer themes for the second probe.

How much answer content a probe extracts also depends on whether the respondents actually provide the information that the probe asks for. Mismatches reduce answer content. We identified two mechanisms that potentially increase the occurrence of mismatches and decrease answer content: the focus of attention and the need for justification. Mismatches only appeared at the specific probe and not at the category-selection probe which is already an indication for the need for justification. For the specific probe, the combined impact of both mechanisms, indeed, led to a strong interaction effect. In the one-screen condition, incidences of mismatches were elevated for the specific probe if respondents received it as the second probe. This might be an indication that respondents pay less attention to the wording of the second probe (outside of the respondents' view on the screen) except if the urge for justification was not satisfied, yet. Only then, some of the respondents move their focus to the verbal information of the second probe. For the two-screen setting, mismatches were lowest if respondents received the specific probe on the second screen and highest if they received it on the first screen. Respondents seem to focus more thoroughly on the task at hand when they receive the specific probe on the second screen. However, if they first receive a specific probe without knowing whether they will get the chance to justify their response, they are more likely to provide a mismatching response. The need for justification seems again to counteract the effect of attention focus.

This pattern of mismatching behavior also has consequences for response quality and respondents' motivation. Concerning response quality, we observed a complex interaction between number of screens and probe order. In addition to the perceived response burden, mismatches push nonresponse behavior. As a consequence, overall nonresponse is lowest if either the specific probe is asked first on one screen or the category-selection probe is asked first on two screens. It is highest when either the category-selection probe is asked first on one screen or the specific probe is asked first on two screens.

Concerning respondents' motivation, we found a significant main effect of probe order but not for number of screens. Respondents that received a category-selection probe as the first probe showed fewer signs of reduced motivation than respondents that received a specific probe as the first probe. It is not surprising that respondents who received a category-selection probe as the second probe overtly complained about receiving a further probe. As already mentioned, respondents often gave mismatching responses at the specific probe when they received the latter as the first probe. That is, the majority of these respondents reported the reasons for choosing an answer value (i.e., the

answer to a category-selection probe) instead of providing a proper answer (i.e., the answer to a specific probe). When these respondents received as their second probe a category-selection probe, they got more easily frustrated since they had already provided the answer to this probe. Indeed, further analysis revealed that two-thirds of respondents that had previously given a mismatching response at the specific probe overtly complained when they received the category-selection as their second probe. However, the number of screens does not seem to impact respondents' motivation much. It is important to note that the overall prevalence of this indicator is very low but in line with the results of Meitinger et al. (2018).

Finally, our last hypothesis predicted a question order effect regarding answer content. We expected a main effect of probe order with respondents experiencing less insecurity about the expected answer scope at the category-selection probe if they receive it after a specific probe. When comparing "it depends" responses for the category-selection probe by experimental group, more respondents provided "it depends" responses if they received the category-selection probe as the first probe, and this question order effect was significant. This is not really surprising because respondents who first received the specific probe could use this probe to narrow down the scope of their response (e.g., define which country they referred to) and apply this scope at the category-selection probe. This is an indication that a context effect appeared at the multiple probes that was—contrary to some other context effects—helpful for the respondent.

Overall, we see that the number of screens and the probe order have an impact on response quality, answer content, and respondents' motivation and this impact is driven by the mechanisms of perceived response burden, focus of attention, need for justification, and question order. The question remains whether we can recommend one specific setting based on our results.

Respondents who received first a category-selection probe on two screens (EG3) outperformed respondents from other settings on most indicators: number of themes, occurrence of mismatches, and overt signs of reduced motivation. In addition, nonresponse levels were also relatively low for this setting. However, respondents in EG3 provided more "it depends" responses than respondents in EG2 or EG4. Therefore, we recommend this implementation (EG3) in all but one scenario: if the goal is to test an item with a key term that has a potentially too large lexical scope, we recommend starting with a specific probe but on the same screen as the category-selection probe (EG2). Why is the one-screen setting preferable in this situation? It is because respondents in EG2 showed less nonresponse, fewer mismatches, and less "it depends" responses than respondents in EG4.

4.1 Limitations

We tested the sequence effect with only one item and the effects found are possibly specific to the item selected. Therefore, replication with different items

and content areas is desirable. However, we built this research on Meitinger et al. (2018) and could replicate several of their findings with regard to question order. In a similar vein, it is important to note that we used a nonprobabilistic sample which is why we cannot make inferences about the German population. Also, though we positioned our experiment near the beginning of the survey, respondents had already received category-selection and specific probes before, which is why carry-over or context effects from these previous probes cannot be fully excluded. Thus, different probe sequences should be assessed in various positions throughout a survey. Finally, this experiment was implemented in a web survey with only desktop respondents. Therefore, we cannot make inferences to smartphone respondents of web surveys. This is particularly the case because the visual design differs for smartphone respondents (e.g., an implementation of multiple probes on the same screen might not be a viable option for these devices).

4.2 Future Research

One interesting extension of this research would be a cross-national implementation. Meitinger et al. (2018) found variations of effect sizes of different indicators across countries. For example, the probe order largely impacted nonresponse in Great Britain and the United States, but in Mexico and Spain, it affected the prevalence of mismatching behavior instead. It would be interesting to assess the combined effect of the number of screens and the probe order in these countries. Unfortunately, our data for this experiment was restricted to Germany. Future research should replicate this experiment in different cultural settings. Another interesting extension of this research would be to use eye-tracking data to disentangle the mechanisms that affect response behavior even further. In particular, regarding the mechanisms of perceived response burden and focus of attention, eye-tracking could provide interesting insights. For example, it would be interesting to disentangle whether respondents do not see relevant information (focus of attention) or whether they register all relevant information but consciously decide against providing all requested information (perceived response burden). Finally, research on optimal implementation of multiple probes should also be extended to mixed device surveys to ensure that also smartphone respondents receive a probe design that optimizes response behavior.

REFERENCES

Beatty, P. C., and G. Willis (2007). "Research Synthesis: The Practice of Cognitive Interviewing," *Public Opinion Quarterly*, 71, 287–311.

Behr, D., W. Bandilla, L. Kaczmirek, and M. Braun (2014). "Cognitive Probes in Web Surveys: On the Effect of Different Text Box Size and Probing Exposure on Response Quality," *Social Science Computer Review*, 32, 524–533.

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. Mannheim, GESIS – Leibniz-Institute for the Social Sciences (GESIS – Survey Guidelines). DOI: 10.15465/gesis-sg_en_023

Behr, D., K. Meitinger, M. Braun, and L. Kaczmirek (2019). "Cross-National Web Probing: An Overview of Its Methodology and Its Use in Cross-National Studies," in *Advances in Questionnaire Design, Development, Evaluation and Testing*, eds. P. C. Beatty, D. Collins, L. Kaye, J. Padilla, G. Willis, and A. Wilmot, pp. 521–544, Hoboken, NJ: Wiley.

Bradburn, N. (1978). "Respondent Burden. Health Survey Research Methods," DHEW Publication No. (PHS) 79, 35–40.

Braun, M., D. Behr, L. Kaczmirek, and W. Bandilla (2015). "Evaluating Cross-National Item Equivalence with Probing Questions in Web Surveys," in *Improving Survey Methods: Lessons from Recent Research*, eds. U. Engel, B. Jann, P. Lynn, A. Scherpenzeel, and P. Sturgis, pp. 184–200, New York: Routledge.

Braun, M., D. Behr, and J. Díez Medrano (2018). "What Do Respondents Mean When They Report to Be 'Citizens of the World'? Using Probing Questions to Elucidate International Differences in Cosmopolitanism," *Quality and Quantity*, 52, 1121–1135.

Braun, M., D. Behr, K. Meitinger, K. Raiber, and L. Repke (2019). "Using Web Probing to Elucidate Respondents' Understanding of 'Minorities' in Cross-Cultural Comparative Research," *ASK: Research and Methods*, 28, 3–20.

Callegaro, M., and C. DiSogra (2008). "Computing Response Metrics for Online Panels," *Public Opinion Quarterly*, 72, 1008–1032.

Christian, L. M., D. A. Dillman, and J. D. Smyth (2007). "Helping Respondents Get It Right the First Time: The Influence of Words, Symbols, and Graphics in Web Surveys," *Public Opinion Quarterly*, 71, 113–125.

Couper, M. P., R. Tourangeau, and K. Kenyon (2004). "Picture This! Exploring Visual Effects in Web Surveys," *Public Opinion Quarterly*, 68, 255–266.

Couper, M. P., R. Tourangeau, F. G. Conrad, and C. Zhang (2013). "The Design of Grids in Web Surveys," *Social Science Computer Review*, 31(3), 322–345.

Dillman, D. A., J. D. Smyth, and L. M. Christian (2009). *Internet, Mail, and Mixed-Mode Surveys: The Tailored Design Method*, Hoboken: Wiley and Sons.

Efremova, M., T. Panyusheva, P. Schmidt, and F. Zercher (2018). "Mixed Methods in Value Research: An Analysis of the Validity of the Russian Version of the Schwartz Value Survey (SVS) Using Cognitive Interviews, Multidimensional Scaling (MDS), and Confirmatory Factor Analysis (CFA)," *Ask: Research and Methods*, 26, 3–30.

Holsti, O. R. (1969). *Content Analysis for the Social Sciences and Humanities*, Reading, MA: Addison-Wesley.

ISSP Research Group. (2015). *International Social Survey Programme: National Identity III—ISSP 2013*. GESIS Data Archive, Cologne. ZA5950 Data file Version 2.0.0. doi:10.4232/1.12312.

Kahneman, D. (1973). *Attention and Effort*, Englewood Cliffs, NJ: Prentice Hall.

Keusch, F. (2014). "The Influence of Answer Box Format on Response Behavior on List- Style Open-Ended Questions," *Journal of Survey Statistics and Methodology*, 2, 305–322.

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

Lee, S., C. McClain, D. Behr, and K. Meitinger (2020). "Exploring Mental Models behind Self-Rated Health and Subjective Life Expectancy through Web Probing," *Field Methods, Online First*, 32, 309–326.

Meitinger, K. (2017). "Necessary but Insufficient: Why Measurement Invariance Tests Need Online Probing as a Complementary Tool," *Public Opinion Quarterly*, 81, 447–472.

———. (2018). "What Does the General National Pride Item Measure? Insights from Web Probing," *International Journal of Comparative Sociology*, 59, 428–450.

Meitinger, K., and D. Behr (2016). "Comparing Cognitive Interviewing and Online Probing: Do They Find Similar Results?," *Field Methods*, 28, 363–380.

Meitinger, K., D. Behr, and M. Braun (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*, 089443931985984.

Meitinger, K., M. Braun, and D. Behr (2018). "Sequence Matters in Online Probing: The Impact of the Order of Probes on Response Quality, Motivation of Respondents, and Answer Content," *Survey Research Methods*, 12, 103–120.

Meitinger, K., and T. Kunz (2018). "Does Quantity Come at the Expense of Quality?—Visual Design Manipulations and Cognition in List-Style Open-Ended Questions in Web Surveys," paper presented at GOR Conference, Cologne, Germany.

Prüfer, P., and M. Rexroth (2005). "Kognitive Interviews [cognitive interviews]," ZUMA How-to-Reihe 15.

Redline, C., D. Dillman, A. Dajani, and M. A. Scaggs (2002). "The Effects of Altering the Design of Branching Instructions on Navigational Performance in Census 2000," *Survey Methodology*, 2.

Schulz, S., K. Meitinger, M. Braun, and D. Behr (2018). "Who's Bad? Eine Analyse Zur Internationalen Vergleichbarkeit Von Maßen Krimineller Einstellungen Mittels Des Web-Probing Ansatzes," in *Kriminologische Welt in Bewegung, Neue Kriminologische Schriftenreihe 117*, eds. K. Boers and M. Schaerff, pp. 406–417, Forum Verlag: Godesberg,

Schwarz, N. (1999). "Self-Reports of Behaviors and Opinions: Cognitive and Communicative Processes," in *Cognition, Aging, and Self-Reports*, eds. N. Schwarz, D. C. Park, B. Knäuper, and S. Sudman, pp. 17–43, Philadelphia: Psychology Press Ltd.

Schwarz, N., and F. Strack (1991). "Context Effects in Attitude Surveys: Applying Cognitive Theory to Social Research," *European Review of Social Psychology*, 2, 31–50.

Schwarz, N, and., H. Bless (1992). "Scandals and the Public's Trust in Politicians: Assimilation and Contrast Effects," *Personality and Social Psychology Bulletin*, 185574–579.10.1177/0146167292185007

Smyth, J. D., D. A. Dillman, and L. M. Christian (2007). "Improving Response Quality in List-Style Open-Ended Questions in Web and Telephone Surveys," paper presented at the Annual Conference of the American Association for Public Opinion Research, Anaheim, CA.

———. (2008). "Context Effects in Internet Surveys," in *Oxford Handbook of Internet Psychology*, Joinson, A. N., McKenna, K. Y. A., Postmes, T., & Reips, U. D. (Eds.), Oxford University Press. pp. 430–445.

Tourangeau, R., L. Rips, and K. Rasinski (2000). *The Psychology of Survey Response*, Cambridge: Cambridge University Press.

Ware, C. (2000). *Information Visualization: Perception for Design*, San Francisco: Morgan Kaufman.

Willis, G. D. (2005). *Cognitive Interviewing: A Tool for Improving Questionnaire Design*, Thousand Oaks, CA: Sage.