Survey Research Methods (2018)
Vol. 12, No. 2, pp. 103-120
doi:10.18148/srm/2018.v12i2.7219

ISSN 1864-3361
http://www.surveymethods.org

Sequence matters in web probing: the impact of the order of probes on response quality, motivation of respondents, and answer content

Katharina Meitinger
GESIS – Leibniz Institute for the Social Sciences
Mannheim, Germany

Michael Braun
GESIS – Leibniz Institute for the Social Sciences
Mannheim, Germany

Dorothée Behr
GESIS – Leibniz Institute for the Social Sciences
Mannheim, Germany

Due to the growing significance of international studies, the need for tools to assess the equivalence of items in international surveys is pressing. Web probing is a powerful tool for identifying the causes of nonequivalence; it incorporates probing techniques from cognitive interviewing into cross-national web surveys. So far, our web probing approach has applied three different probe types – category-selection probes, specific probes, and comprehension probes – to inquire about different aspects of an item. Previous research has mostly asked one probe type per item, but in some situations it might be preferable to assess potentially troublesome items with multiple probe types. However, empirical evidence is missing on whether the sequence of probe types has an impact on response quality, respondents' motivation, and answer content. In this study, we report evidence from a web experiment that was conducted with 1,354 respondents from Germany, Great Britain, the U.S., Spain, and Mexico in June 2014. In this experiment, we asked respondents three different probes for one item, and we manipulated the sequence of probes in each experimental condition. Our research indicates that the sequence in which different probe types are asked has an impact on response quality, the respondents' motivation, and probe answer content. However, the respondents in the five countries reacted differently to the variation in the probe sequence, suggesting that response behavior to probes is partly culturally driven.

Keywords: web probing; probes; cross-cultural; order of probes

1 Introduction

With the large increase in cross-national data production in social science research (Harkness, 2008; Smith, 2010) comes the challenge of adequately assessing the equivalence of items in international surveys before drawing substantive conclusions; after all, different types of bias (e.g., construct bias, sample bias or item bias; see Van de Vijver & Leung, 2011; Van de Vijver & Poortinga, 1997) can lead to a systematic under- or over-estimation of differences across groups (Davidov, Meuleman, Cieciuch, Schmidt, & Billiet, 2014). Quantitative approaches usually assess equivalence (or the lack thereof) by applying measurement invariance tests that use multigroup confirmatory factor analysis (MGCFA) (Jöreskog, 1971), alignment (Asparouhov & Muthén, 2014), exploratory structural equation modeling (Asparouhov & Muthén, 2009) or Bayesian structural equation modeling (BSEM; see Muthén & Asparouhov, 2012); for an overview, see Davidov et al. (2014). Quantitative approaches such as the multiple indicators multiple causes (MIMIC) model (Davidov et al., 2014) or the multilevel structural equation models (MLSEMs) (Davidov et al., 2016; Davidov et al., 2014; Jak, Oort, & Dolan, 2013, 2014; Meuleman & Schlüter, 2018) aim to explain missing comparability by controlling for differential item functioning on the micro level or by introducing conceptual predictor variables on the macro level (see Davidov et al., 2014 and Meitinger, 2017 for a more detailed discussion). In addition to quantitative approaches, a variety of qualitative methods exist that can provide insights into the equivalence of measures and the reasons for nonequivalence, notably cross-cultural cognitive interviewing (CCCI; e.g., Fitzgerald, Widdop, Gray, & Collins, 2009; Goerman & Caspar, 2010; K. Miller, Mont, Maitland, Altman, & Madans, 2011; for a research synthesis on CCCI see Willis, 2015; and web probing Behr, Meitinger, Braun, & Kaczmirek, 2017; Braun, Behr, Kaczmirek, & Bandilla, 2014). Web probing is a powerful tool for iden-

Contact information: Katharina Meitinger, GESIS – Leibniz Institute for the Social Sciences, PO Box 122155, 68702 Mannheim (e-mail: katharina.meitinger@gesis.org)

tifying the causes of nonequivalence; it involves incorporating probing techniques from cognitive interviewing in cross-national web surveys. Web probing is particularly useful for detecting cases of construct bias and item bias. Construct bias means that the construct measured is not identical across cultures (Van de Vijver & Poortinga, 1997), and items might be biased due to ambiguous source items, poor item translation, inapplicability of item contents or different connotations associated with the item wording in some countries (He & van de Vijver, 2012; Van de Vijver & Leung, 2011).

Probes are follow-up questions that ask respondents to provide additional information about a survey item (Beatty & Willis, 2007). In our web probing studies, the respondents typically receive a probe on a separate screen directly after responding to the item that needs to be tested (Braun et al., 2014). Different probe types can address different aspects of an item. Our previous web probing studies applied three probe types: 1) A category-selection probe asking respondents for the reasons why a certain answer category has been chosen; 2) A specific probe encouraging respondents to provide additional information on a particular detail of the item; 3) A comprehension probe requesting a definition of a specific term (Prüfer & Rexroth, 2005; Willis, 2005).

Previous web probing has mostly asked one probe per item, but in some situations it might be preferable to assess potentially troublesome items with multiple probes. This could be due to the following reasons: If questionnaire designers are uncertain which aspect of an item might be problematic, they can ask several probes that address the different potentially problematic aspects of an item. Furthermore, problematic issues of an item might be located at different stages of the question-answer process. Following Tourangeau et al.'s approach, the respondents have to first comprehend the question text, then they have to retrieve the relevant information from their memory, arrive at a judgment, and finally report their answer selection (Tourangeau, Rips, & Rasinski, 2003). Different probe types address different stages in this question-answer process. For example, the comprehension probe addresses issues that are related to the comprehension stage, while a category-selection probe aims at finding problems related to the response selection, among others (Collins, 2014). Additionally, in cross-national research different aspects of an item or different response stages might be problematic in different countries, which is also reflected in the fact that cross-cultural cognitive interviews tend to apply a wide variety of probe types (Willis, 2015).

In the web mode, different probes need to be decided on and programmed in advance because web probing lacks the interactivity of traditional cognitive interviewing (Meitinger & Behr, 2016) where the interviewer can ask spontaneous and emergent probes (Willis, 2005) that are adapted to the interview situation at any time.

Despite the potential benefits of asking multiple probes, empirical evidence is missing with regard to asking multiple probes in web surveys and whether this has an impact on response quality, the respondents' motivation, and answer content. Two aspects that might influence response behavior with open-ended questions need to be considered in this context: increasing response burden and the impact of probe sequence.

1.1 Increasing Response Burden

Since respondents must write their answer instead of simply choosing an answer option (Keusch, 2014), open-ended questions, such as probes, impose a higher response burden on respondents (Bradburn, 1978) and are therefore potentially more affected by issues of response quality (e.g., higher item nonresponse Barrios, Villarroya, Borrego, & Ollé, 2011) than closed items in web surveys. Furthermore, in web surveys there is no interviewer who could provide a motivation for answering these "burdensome" questions (Meitinger & Behr, 2016). By asking multiple probes, the imposed response burden further increases, which might tempt respondents to write shorter responses for the second or third probe – which could reduce answer content – or, worse, to opt for a probe nonresponse altogether.

1.2 Impact of Probe Sequence

In addition to the increased response burden, the sequence in which the probes are asked might also have an impact on response behavior. For example, is it preferable to first ask a category-selection probe and then follow up with a specific and comprehension probe? Or would we facilitate the respondents' task by first asking a comprehension probe and then following up with a specific and category-selection probe? The question of the optimal probe sequence is not trivial, given that previous research reported on incidences of mismatching probe responses (Behr, Bandilla, Kaczmirek, & Braun, 2014; Meitinger & Behr, 2016). A mismatch occurs, for example, when a respondent gives an answer to a category-selection probe (e.g., explains the reasons for answer selection) in response to a comprehension probe. This may be due to respondents having been exposed and habituated to category-selection probes earlier on in the survey (Behr, Bandilla, et al., 2014) or to respondents simply taking the probe as encouragement to elaborate on their opinion (instead of trying to figure out the definition of a term used in the item), which may be appreciated by some respondents (Couper, 2013). In addition to the obvious effect that mismatching responses increase the percentage of non-substantive responses, mismatching responding might also reduce the motivation of respondents to provide a probe answer for subsequent probes (e.g., respondents who explained the reasons for their answer selection at a comprehension probe might be unmotivated to repeat these reasons at the

actual category-selection probe when this probe is asked in second position). Therefore, it is of great importance to find an ideal probe sequence that avoids the occurrence of mismatches and keeps the response burden as low as possible.

2 Research Goals

In this article, we aim to answer the following research questions:

1. Is the quality of probe responses affected by the sequence in which the different probe types are asked in relation to a single item? We gauged response quality by the length of probe responses, incidences of probe nonresponse, and incidences of probe mismatches.
2. Does the sequence of probes have an impact on the motivation of respondents? In particular, do respondents show overt signs of reduced motivation? We measure reduced motivation with explicit utterances in this regard.
3. Does the sequence of probes have an impact on the answer content? In this context, we use the number of mentioned themes per respondent as a proxy for answer content.

For all research questions, we also evaluate whether the probe sequence has a differential impact on respondents from different countries.

3 Methods and Data

3.1 Participants

This study uses data from a survey conducted with 2,685 respondents in June 2014. The survey participants from Germany, Great Britain, the U.S., Spain, and Mexico were drawn from non-probability online panels based on quotas for age (18-30, 31-50, and 61-65), gender, and education (lower and higher; see Table A1 in the Appendix for a detailed description of the distinction between low and high education). All quotas were met. The main panel provider was respondi[1], 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 panel provider cooperated with international partners, the choice of which we could not control. Since we were interested in open-ended answers and wanted to avoid confounding effects due to differences in devices used, we ensured that respondents entering the survey with a small online device (e.g., a smartphone) could not participate in our survey. Having programmed the survey ourselves, the visual presentation of the survey was identical for all countries. Half of the respondents were randomly selected for this experiment, which is a total of 1,354 respondents. Table 1 summarizes quota characteristics for the respondents in the experiments by country.

Table 1
Mean Age, Gender Distribution, and Percentage of High Education for Respondents in the Experimental Sample by Country and Total

	Age (Mean)	Women %	H. Educ. %
Germany	42	52	53
Great Britain	42	53	51
United States	40	52	50
Spain	41	54	48
Mexico	41	49	53
Total	41	52	51

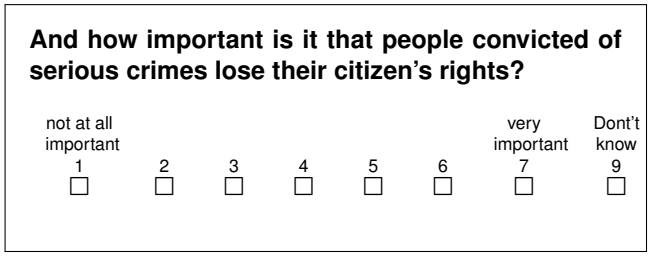

Figure 1. The closed item

3.2 Tested Item

The tested item is part of the "people's rights in a democracy" battery from the 2014 ISSP module on Citizenship (ISSP 2016). The item battery was introduced by the following statement: "There are different opinions about people's rights in a democracy. On a scale of 1 to 7, where 1 is not at all important and 7 is very important, how important is it. . ." The closed item that our research pertains to was: "And how important is it that people convicted of serious crimes lose their citizen rights?" Respondents had to choose their answer on a seven-point scale running from "not at all important" to "very important." A "can't choose" option was also offered (see Figure 1). We used the original ISSP translations for the five countries. Table A2 in the Appendix contains all ISSP item translations[2].

The item was selected because we expected several details of the question wording to be problematic for the cross-national comparability of the data. First, respondents might differ in the definition of what constitutes a "serious crime."

[1] https://www.respondi.com/EN/

[2] Although we used the original ISSP translations, there is always the risk that the item itself is insufficiently comparable. However, the substantive responses to the probes indicate that the item is sufficiently understood in all countries. In addition, any problematic item translation should appear in both experimental groups to a similar degree; therefore the question of comparability is should not affect our experimental manipulation.

Figure 2. Sequence of probes in experimental group 1 (Categ.)

Figure 3. Sequence of probes in experimental group 2 (Compr.)

In addition to country-specific demarcation lines for "seriousness," the translation of "crime" itself might also play a role. Davidov et al. (2014) give the example of the Danish translation of the term "crime" in a European Social Survey item, which carried unintended associations and thereby triggered item bias. Second, the term "citizen rights" is rather vaguely formulated, and some respondents might struggle to understand it. Third, respondents in different countries might apply different reasoning for choosing a certain answer category. Finally, even the direction of the item might be unclear, that is, whether democratic rights entail that people convicted of serious crimes lose or do not lose their citizenship rights.

3.3 Procedure

The item battery was part of a longer questionnaire replicating questions from the ISSP modules on Citizenship as well as Family and Gender Roles. The experiment was implemented near the beginning of the questionnaire (respondents answered two probes before they received this experiment).

All respondents first answered the closed item on one screen and subsequently received three probes administered on separate screens. The two experimental conditions were assigned at random to participants in each country sample, forming two experimental groups per country. Group 1 ("Categ.") first received a category-selection probe asking for the reasons why a certain answer category had been chosen, followed by a specific probe ("What particular citizen rights did you have in mind when you were answering the question?"), and then a comprehension probe ("What do you consider to be a 'serious crime?'"). Group 2 ("Compr.") first received the comprehension probe, which was followed by the specific probe, and finally, the category-selection probe (see Figures 2 & 3). To ensure optimal comparability of the probes themselves, we employed team-based translation (Harkness, 2003) for the probe translation: The probes for each language were translated by two professional translators who had the same cultural background as the countries in our study. Subsequently, the research team convened with the translators to discuss and arrive at final probe versions. Table A3 in the Appendix contains the probe translations for each country.

Figure 4 shows illustrations of the probes that were implemented in the web survey. Based on previous research on the optimal text-box design of probe questions in web surveys (Behr, Bandilla, et al., 2014), we selected a small text box for the specific probe and large text boxes for the category-selection and comprehension probes. This text-box design helps the respondents to decide which type of answer is expected (e.g., the small text box indicates that a short answer, possibly including only a few key words, is expected). Based on the probe answers, a separate coding schema was developed for each of the probes. All probe responses were coded, and a sample of 20% of the probe answers was coded a second time to assess intercoder reliabilities (category-selection probe: 97%; specific probe: 91%; comprehension probe: 92%). The coding team discussed and corrected all instances of deviating coding.

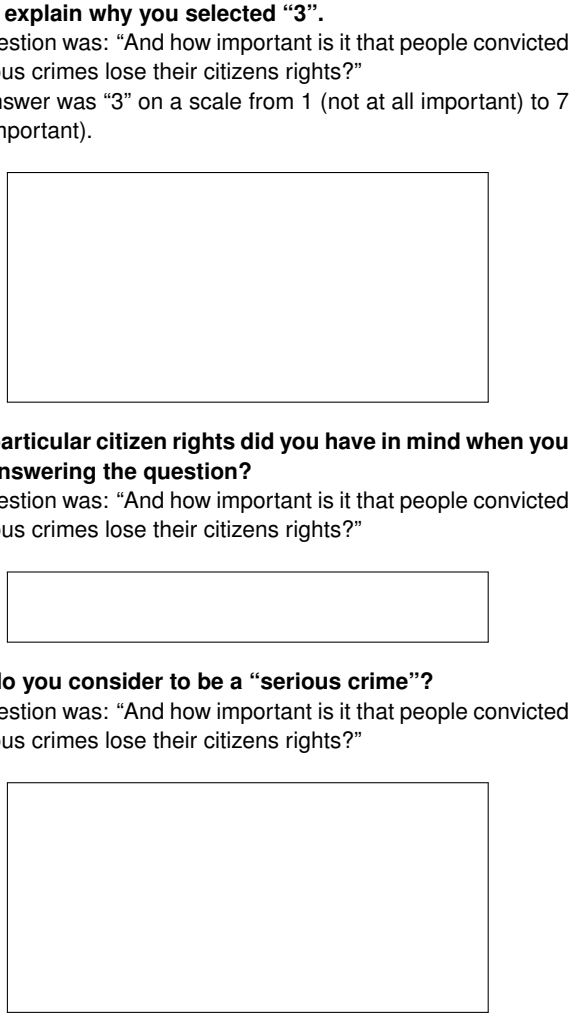

Figure 4. Category-selection probe, specific probe, and comprehension probe (sequence of Group 1)

3.4 Analysis

We gauged the quality of probe responses with three indicators: the length of the probe responses, the incidences of probe nonresponse, and the proportion of probe mismatches.

1. The length of the probe response is the most commonly used proxy for response quality in research about open-ended questions, with longer responses usually being interpreted as "better" responses.[3] We assessed the length of the probe responses by the number of characters.

2. We also gauged the response quality by the proportion of respondents giving a probe nonresponse. Previous studies investigating nonresponse to open-ended questions defined non-respondents as any respondents submitting a completely blank text box (one character would already qualify as a substantive response; e.g., A. L. Miller & Lambert, 2014). However, respondents can also give a probe nonresponse by typing several characters. Therefore, we coded all probe answers as probe nonresponses if they had no text entry, had unintelligible letter combinations (e.g., "xcvbnm"), contained refusals (e.g., "n/a," "no comment"), were "don't knows," and were meaningless or incomprehensible answers (e.g., "just cause," "awesome") (see Behr, Braun, Kaczmirek, & Bandilla, 2014; for a similar approach Holland & Christian, 2009).

3. Our third indicator for response quality was the proportion of mismatching probe responses. A mismatching probe response occurs when respondents write an answer to a different probe type than required. For example, respondents explain their reasons for choosing a certain answer value at a comprehension probe even though here are they supposed to define a key term in their own words (e.g., a respondent explains why s/he selected answer value number "7" instead of describing what for him/her constitutes a "serious crime").

Binary codes assessing the presence of mismatching responses (0: "no mismatch"; 1: "mismatch") were assigned to all probe responses differentiating between probe types (e.g., for the category-selection probe we applied one binary code for respondents actually answering a specific probe and one binary code for respondents answering a comprehension probe).

Our second research goal was to assess whether respondents' motivation varied across experimental groups. We used the proportion of respondents' overt signs of reduced motivation as an indicator, given that several respondents expressed discontent or increasing frustration with the probes. We coded all probe responses where respondents complained that they had already answered the question (e.g., British respondent: "I HAVE ALREADY ANSWERED THIS!!!!") as indicating respondents' reduced motivation.

Finally, our third research question was whether the sequence of probes had an impact on the answer content. To answer this research question, we developed a separate coding schema[4] for the responses to each probe type that captured various substantive themes (CSP: reasons for choosing an answer category; SP: different types of citizen rights; COP: definitions of serious crimes) mentioned by the respondents. For this article, we use the number of mentioned themes per respondent as a proxy for answer content.

4 Results

4.1 Is the Quality of Probe Responses Affected by the Sequence in which the Different Probe Types are Asked?

Length of probing answers. Table 2 shows the mean number of characters combined for all three probes by experimental group and for each country. The table also presents the results separately for all respondents (substantive respondents and nonrespondents) and for substantive respondents only (respondents who gave a substantive response to all three probes). It also contains for each country the results of a two-sample *t-test*[5]. To begin with, the results in this table show that respondents from the five countries differ in their response length. Mexican and Spanish respondents give longer responses than respondents from the other countries. Does this mean that Spanish-speaking respondents write answers with a higher response quality than respondents from other countries? Not necessarily, since the response length may also differ due to linguistic reasons or varying communication styles, and we cannot disregard these reasons as an alternative explanation. As a consequence, response length should not be used as an overall indicator for response quality but should only be used to compare experimental groups within countries. Therefore, we refrain from evaluating response quality across countries by length of response.[6]

When comparing between the experimental groups we cannot find a clear direction of the effect of the probe sequence. Respondents who started with a comprehension probe wrote slightly longer responses in Germany, Great Britain, and Mexico than respondents who first received a category-selection probe. However, we cannot reject the null

[3] Although longer responses are usually regarded as better responses, current research has started to question this assumption because response length might reduce intercoder reliability (Andrews, 2005; Conrad, Couper, & Sakshaug, 2016; Denscombe, 2008).

[4] The full coding schemas are available from the authors upon request.

[5] Since our data did initially not meet the criteria of a normal distribution, we replaced the values of extreme outliers (above/below highest and lowest percentile) with the values of the highest and lowest percentile of respondents.

[6] In general, the common assumption that long answers automatically indicate high quality and short answers low quality should be critically evaluated in the research community. Depending on how the answers are used, short answers may provide equally good content. This is a matter for further research.

hypothesis that the length of responses is equal in both experimental groups. In contrast, respondents in Spain and the U.S. wrote longer responses when starting with a category-selection probe; the effect was significant in the U.S. (Categ.: Mean = 171, Std. Dev. = 135; Compr.: Mean = 139, Std. Dev. = 116), but it had only a small effect size [$t(278) = 2.13$, $p = 0.03$, $d = -0.25$], and no significant effect could be found for the subsample of substantive respondents. Therefore, no clear sequence effect can be detected with regard to response length.

Probe nonresponse. As seen in Table 3, respondents' nonresponse rates differ across countries. In general, respondents from Mexico and Spain less often gave a nonresponse than respondents from Germany, Great Britain, and the U.S. More importantly, though, the sequence in which the different probes were asked had an impact on probe nonresponse in some countries. On average, probe nonresponse was lower for all probe types in experimental Group 1 (Categ.), and these effects were highly significant in the *Fisher's exact test* for the category-selection probe (FET, $p = 0.00$) and specific probe (Mean, $p = 0.01$). Interestingly, the impact of probe sequence on probe nonresponse varied across countries: U.S. respondents were particularly affected since a higher proportion of respondents in experimental Group 2 (Compr.), which had been given a comprehension probe first, gave a probe nonresponse both at the category-selection probe and the specific probe compared to experimental Group 1 (Categ.) and these differences were highly significant (U.S. category-selection: FET, $p = 0.00$, OR= 3.01; specific: FET, $p = 0.02$, OR= 2.51). The odds ratio for the category-selection probe, for example, was 3.01, indicating that U.S. respondents had a three times higher chance of writing a nonresponse at the category-selection probe when they received this probe as the third probe (Compr.).

The same pattern emerged for British respondents, and the differences between the experimental groups were also statistically significant for the category-selection and specific probes (GB category-selection: FET, $p = 0.02$, OR= 2.45; specific: FET, $p = 0.05$, OR= 2.58). In contrast, probe nonresponse in Mexico and Spain differed only slightly across experimental groups. Nearly no impact and no clear pattern could be found for the German experimental groups. Thus, the impact of probe sequence on probe nonresponse differed across countries.

Mismatching probes. Table 4 summarizes the occurrence of mismatching probe responses by probe type, experimental group, and country. When a category-selection probe was asked, respondents in all countries rarely responded in a way that would be typical for a specific or comprehension probe. Mismatches appeared more frequently when respondents had to answer a specific probe. However, most cases of mismatching probe responses occurred when a comprehension probe was asked as the first of multiple probes (columns on the right, Table 4). On average, more than 20% of respondents in experimental Group 2 (Compr.) answered the comprehension probe, which was given as the first probe, with a response to a category-selection probe (that is, they provided a reason for selecting a specific answer value instead of providing a definition of what constitutes a "serious crime"). However, the respondents in experimental Group 1 (Categ.) rarely provided a mismatching response when they had to provide an answer to a comprehension probe (which came in third position in Group 1). The two-sided *Fisher's exact test* comparing the experimental groups for each probe and country separately indicates a highly significant difference between the two experimental groups and this is the case for each country (see Appendix Tables B1–B3 for detailed results of the *Fisher's exact test*). This is a clear indication that response quality can be improved by first asking respondents a category-selection probe instead of a comprehension probe when multiple probes are asked.

Although the sequence effect appeared in all countries, the strength of this effect differed across the five countries. German respondents provided the fewest mismatches (12.95%). British and American respondents tended to give slightly more mismatches (17.33% and 17.73%). Contrary to the results with regard to probe nonresponse, Spanish and Mexican responses were highly affected by mismatching probe responses (22.76% and 32.56%) when respondents first received a comprehension probe. We do not have a ready explanation of these country differences in the effect of probe sequence on the incidence of mismatches. This issue should be further investigated in future research.

Overall, there are clear indications that the sequence in which category-selection and comprehension probes are asked has an impact on response quality. Respondents seem to clearly prefer to first answer category-selection probes; otherwise the response quality might be lowered by increased probe nonresponse rates and incidences of probe mismatches. However, the sequence effect does not show up in the same way for the three indicators we have used for response quality. While for the length of probing answers there is practically no effect, and for probe nonresponse the effect is significant only in some of the countries in the study, mismatches are considerable and present in all the countries, though to different degrees.

4.2 Does the Motivation of Respondents Differ Across Split Conditions?

Responding to multiple probes creates an increased cognitive burden for respondents. Since there is no interviewer present who could exert a motivating effect on web survey respondents (in comparison to cognitive interviewing), multiple probes might reduce respondent motivation to answer such open-ended questions. We assessed the motivation to answer multiple probes with the percentage of re-

Table 2
Mean Length, Standard Deviation, and Two-sided T-test of Probe Responses in Characters by Experimental Group and Country for All and Substantive Respondents

		Categ.[a]		Compr.[b]		t-test		
Probe Split	N	Mean	Std. Dev.	Mean	Std. Dev.	df	t	p
All respondents (N = 1,354)								
Germany	268	154	110	181	125	266	−1.87	0.06
GB	277	175	146	181	141	275	−0.36	0.72
U.S.	280	171	135	139	116	278	2.13	0.03
Spain	252	251	170	233	133	250	0.93	0.35
Mexico	277	256	148	261	145	275	−0.28	0.78
Substantive respondents (N = 859)								
Germany	174	174	117	205	140	172	−1.59	0.11
GB	172	182	145	207	148	170	−1.12	0.26
U.S.	161	192	135	189	123	159	0.13	0.90
Spain	166	263	158	244	131	164	0.80	0.42
Mexico	186	266	154	270	153	184	−0.15	0.88

[a] First probe: category-selection [b] First probe: comprehension

Table 3
Probe Nonresponse by Probe Type, Country, and Experimental Group in Percent and Results for Two-sided Fisher's Exact Test

Probe		Category-selection			Specific			Comprehension		
		Categ.[a,c]	Compr.[b]		Categ.[a]	Compr.[b]		Categ.[a]	Compr.[b,c]	
Split	N	%	%	p	%	%	p	%	%	p
Germany	268	10.85	10.79	1.00	10.85	9.35	0.69	3.88	4.32	1.00
GB	277	7.87	17.33	0.02	4.72	11.33	0.05	4.72	5.33	1.00
U.S.	280	7.91	20.57	0.00	7.91	17.73	0.02	5.76	9.22	0.37
Spain	252	3.10	5.69	0.37	4.65	7.32	0.43	4.65	3.25	0.75
Mexico	277	1.35	5.43	0.09	4.05	3.88	1.00	2.03	3.88	0.48
Total	1,354	6.10	12.32	0.00	6.40	10.12	0.01	4.17	5.28	0.37

[a] First probe: category-selection [b] First probe: comprehension
[c] Groups receiving the respective probe in first position

spondents' statements of reduced motivation.

Respondents' overt signs of reduced motivation. Table 5 shows the percentage of respondents who openly complained in their response that they had already answered the question. As mentioned above, open complaints in web surveys can be seen as an indicator of severe frustration of the respondents. Respondents in all countries complained the most when they received the category-selection probe as their third probe (Group 2: Compr.). In contrast, respondents who first received the category-selection probe (Group 1: Categ.) did not show any indications of reduced motivation at the comprehension probe, which was their third probe. The difference between the experimental groups is statistically significant for all countries (two-sided *Fisher's exact test*) except for the U.S. This finding is in line with the previous results that respondents clearly prefer to first answer a category-selection probe. Interestingly, the percentage of complaining German respondents was comparatively high. Thus, the German results were, for the first time in these experiments, "outstanding."

It is not surprising that respondents who overtly complained about receiving a further probe mostly received a category-selection probe as their third probe. As already mentioned, a surprising number of respondents gave mismatching responses at the comprehension probe when they received the comprehension probe as their first probe. That is, the majority of these respondents wrote the reasons for choosing an answer value (that is, the answer to a category-selection probe) instead of providing a definition of the term "serious crime" (that is, the answer to a comprehen-

Table 4
Mismatching Probe Responses by Probe Type, Experimental Group, and Country in Percent and Results for Two-sided Fisher's Exact test

Probe:		Category-selection				Specific				Comprehension			
Mismatch:		Specific		Comprehension		Categ. select.		Comprehension		Categ. select.		Specific	
Split	N	Categ.[a] %	Compr.[b] %	Categ.[a] %	Compr.[b] %	Categ.[a] %	Compr.[b] %	Categ.[a] %	Compr.[b] %	Categ.[a] %	Compr.[b] %	Categ.[a] %	Compr.[b] %
Germany	268	0	0	1	0	5	4	1	1	2	13**	0	1
GB	277	0	1	0	0	3	7	6	2	2	17**	0	1
U.S.	280	0	4*	0	0	5	5	5	4	1	18**	1	1
Spain	252	2	0	1	1	4	4	4	4	2	23**	0	0
Mexico	277	0	2	0	2	4	9	4	3	3	33**	0	0
Total	1,354	0	1*	0	1	4	6	4	3	2	20**	0	0

[a] First probe: category-selection [b] First probe: comprehension.
* $p \leq 0.05$ ** $p \leq 0.01$

Table 5
Indications of Reduced Motivation in Percent: Respondents' Statements. Two-sided Fisher's Exact Test

Probe		Category-selection			Specific			Comprehension		
		Categ.[a,c]	Compr.[b]		Categ.[a]	Compr.[b]		Categ.[a]	Compr.[b,c]	
Split	N	%	%	p	%	%	p	%	%	p
Germany	268	0	6	0.00	0	0	-	0	0	-
GB	277	0	4	0.03	0	0	-	0	0	-
U.S.	280	0	3	0.12	0	0	-	0	0	-
Spain	252	0	7	0.00	0	2	0.24	0	0	-
Mexico	277	0	3	0.05	1	1	1.00	1	0	1.00
Total	1,354	0	5	0.00	0	0	0.62	0	0	0.50

[a] First probe: category-selection [b] First probe: comprehension [c] Groups receiving the respective probe in first position

sion probe). When these respondents received a category-selection probe as their third probe, they got more easily frustrated since they thought that they already provided the answer to this probe. Indeed, further analysis revealed that 45% of respondents who had previously given a mismatching response to the comprehension probe overtly complained when they received the category-selection as their third probe.

4.3 Does the Sequence of Probes Have an Impact on the Answer Content?

In addition to triggering variations in response quality and respondent motivation, the sequence of probes can also impact the answer content, which we define as the number of themes that respondents mention. Table 6 summarizes the average number of themes mentioned by respondents by country, probe type, and split condition. Only those respondents who provided a substantive response are considered.[7]

On average, respondents thought about the fewest themes (reasons for answer selection) at the category-selection probe and the most themes at the comprehension probe (examples for serious crimes). Due to the different response tasks involved in these different probe types, this pattern is not surprising. It is easier to think about – or retrieve – several examples of serious crimes (comprehension probe) or citizens' rights (specific probe) than reporting multiple reasons for an answer selection. Once again, respondents in the five countries differed in their response behavior. On average, German and U.S. respondents mentioned the fewest themes, British respondents referred to slightly more themes, and Spanish and Mexican respondents were in the lead with regard to the number of themes.

However, we could not find any clear pattern with regard to probe sequence. For the category-selection and the specific probe, no clear pattern emerged, and we could not reject the null hypothesis that the number of themes is equal in both experimental groups in all countries. At the comprehension probe, the probe sequence did not have a clear impact on the number of mentioned themes either. In Germany and Great Britain, respondents mentioned more themes at the comprehension probe when they received it as their first probe. In contrast, respondents from the U.S., Spain, and Mexico were more productive at this probe when they received the comprehension probe in third position. However, only the mean differences for Germany and Mexico were statistically significant, but they had only a small effect size (Germany: $t(229) = -2.16$, $p = 0.03$ (two-tailed), $d = -0.28$; Mexico: $t(216) = 2.55$, $p = .01$ (two-tailed), $d = 0.36$).

All in all, the probe sequence did seem to have an impact on the number of mentioned themes. However, it is important to note that these numbers are calculated on the basis of substantive responses. Given the elevated probe nonresponse in certain countries, it might be that the mean number of mentioned themes is also indirectly affected by the sequence effect on probe nonresponse.

5 Discussion & Limitations

We set out to explore the effect of probe sequence (category-selection probe at the first vs. the third position) on probe response quality, respondents' motivation, and probe answer content. For probe response quality, we used three indicators: length of the answer provided by respondents, probe nonresponse, and the amount of mismatching probe answers (i.e., respondents gave an answer not corresponding to the actual probe type, e.g., they wrote about their reasons for selecting a specific answer value even though they had been asked a comprehension probe). The length of the answers was virtually unaffected by probe sequence while

[7] Since our data did not initially meet the criteria of a normal distribution, we replaced the values of extreme outliers (above/below highest and lowest percentile) with the values of the highest and lowest percentile of respondents.

Table 6
Number of Mentioned Themes per Respondent and T-test of Probe Responses by Experimental Group and Country for Substantive Respondents

	Categ.[a]		Compr.[b]		t-test		
	Mean	Std. Dev.	Mean	Std. Dev.	t	df	p
Probe: Category-selection							
Germany	1.1	0.3	1.0	0.2	0.82	214	0.41
GB	1.1	0.3	1.2	0.4	−0.98	228	0.33
U.S.	1.1	0.4	1.2	0.4	−1.21	221	0.23
Spain	1.2	0.4	1.1	0.3	1.87	220	0.06
Mexico	1.1	0.4	1.1	0.3	1.11	258	0.27
Total	1.1	0.4	1.1	0.3	0.66	1149	0.51
Probe: Specific							
Germany	1.4	0.7	1.4	0.8	−0.28	220	0.78
GB	1.6	0.9	1.6	0.8	0.23	216	0.82
U.S.	1.4	0.6	1.3	0.6	0.76	213	0.45
Spain	1.7	0.9	1.6	0.8	0.62	208	0.54
Mexico	1.7	1.9	1.6	0.8	0.77	232	0.44
Total	1.6	0.9	1.5	0.8	0.04	1097	0.30
Probe: Comprehension							
Germany	2.2	1.0	2.5	1.1	−2.2	229	0.03
GB	2.4	1.4	2.7	1.3	−1.4	231	0.16
U.S.	2.1	1.2	2.1	1.1	0.4	229	0.67
Spain	2.7	1.2	2.5	1.1	1.4	205	0.17
Mexico	3.0	1.3	3.0	1.2	2.6	216	0.01
Total	2.5	1.3	2.5	1.2	0.6	1118	0.52

[a] First probe: category-selection [b] First probe: comprehension

probe nonresponse increased for the category-selection and specific probes when respondents first received a comprehension probe; however, this effect was only present in two of the countries in our study, namely, Great Britain and the U.S. The percentage of mismatching answers at the comprehension probe was significantly higher in all countries when the comprehension probe was asked as the first probe. This effect was strongest in Mexico and weakest in Germany. With regard to respondents' motivation, which was measured by complaints at the third probe, we also found significant differences between split conditions in all countries: such complaints only appeared when the category-selection probe was asked as the third probe but not when it was asked as the first probe. Finally, with regard to answer content, we did not find any statistically significant differences between the experimental groups in any country for the category-selection or the specific probe, and for the comprehension probe we only found a significant difference for Germany and Mexico.

5.1 Limitations

We tested the sequence effect with only one item, and the found effects are not generalizable but specific to the experiment; therefore, replication with different items and content areas is desirable. In addition, we found several variations in response behavior to open-ended questions across countries; these require further research to fully understand. Replication of this study with the same but also with different countries is necessary to fully assess whether these cross-cultural variations are stable across studies. Although we took care that the visual presentation was identical across countries, and our panel provider adhered to high quality criteria (ISO 26362), we cannot fully exclude differences in the recruitment of respondents in the five countries. In a similar vein, it is important to note that we used a non-probabilistic sample; therefore, we cannot make inferences about the countries' populations. Additionally, since the countries in our study were not randomly selected but purposively chosen, our study is an "accumulation" of results obtained in the five countries and can, thus, not be generalized. We also cannot fully exclude that the found variations across countries

might be due to cross-cultural differences in the understanding of the item. Although a content analysis of the mentioned themes (Schulz, Meitinger, Braun, & Behr, forthcoming) revealed that respondents in the five countries have similar associations regarding which crimes constitute a "serious crime," future research should replicate this experiment with bilinguals within a country to fully disentangle country and language effects. Finally, though we positioned our experiment near the beginning of the survey, respondents had already received category-selection and specific probes before the experiment, which is why carry-over or context effects from these cannot be fully excluded. Thus, different probe sequences should be assessed in various positions throughout a survey.

5.2 Recommendations & Future Research

Given our findings, we recommend to first ask a category-selection probe and then follow up with a specific and a comprehension probe when asking multiple probes. When respondents first received a category-selection probe, the number of mismatches was clearly lower in all countries and respondents were more motivated. The indicators "length of the answer" and "number of themes" are clearly weaker and are not unambiguous indicators: After all, in the end it is not clear that longer answers and more themes mentioned are necessarily and in every case better per se.

A second aspect that the experiment could reveal is the culture-specific response behavior of the respondents in the five countries. British and U.S. respondents were more prone to probe nonresponse, whereas Spanish and Mexican responses were more often affected by mismatching responses when the comprehension probe appeared in the first position. German respondents were clearly less affected by the changing probe sequence. These results suggest that there are differences in the degree to which and the way respondents from different countries react to questionnaire design variations, which might be related to cultural discourse norms. British and U.S. respondents did not react by giving a mismatching answer but by an alternative behavior – nonresponse – to a probe type they did not like. Further research should aim to uncover the mechanisms that drive these diverging response behaviors since differences in how respondents react to variations in the questionnaire design might have implications for the universal applicability of survey methodological research in this area. The majority of methodological studies on surveys are conducted in the U.S. context. The question remains whether all findings regarding questionnaire design are automatically transferable to all countries or whether we must adapt our strategies to the specific cultural settings.

Acknowledgments

The authors thank the anonymous reviewers for their helpful comments on earlier versions of this manuscript. This research was funded 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). An earlier version of this paper was presented at the Second International Conference on Survey Methods in Multinational, Multiregional, and Multicultural Contexts (3MC) in Chicago.

References

Andrews, M. (2005). *Who is being heard? Response bias in open-ended responses in a large government employee survey.* 60th Annual Conference of the American Association for Public Opinion Research, Miami Beach, FL.

Asparouhov, T. & Muthén, B. (2009). Exploratory structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, *16*(3), 397–438.

Asparouhov, T. & Muthén, B. (2014). Multiple-group factor analysis alignment. *Structural Equation Modeling: A Multidisciplinary Journal*, *21*(4), 495–508.

Barrios, M., Villarroya, A., Borrego, A., & Ollé, C. (2011). Response rates and data quality in web and mail surveys administered to PhD holders. *Social Science Computer Review*, *29*(2), 208–220.

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

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.

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.

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 items. GESIS – Survey Guidelines. Mannheim: GESIS – Leibniz-Institute for the Social Sciences. doi:10.15465/gesis-sg_en_023

Bradburn, N. (1978). Respondent burden. *Health Survey Research Methods, DHEW Publication No.(PHS)*, *79*(3207), 35–40.

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. Scherpenzeel, & P. Sturgis (Eds.), *Improving survey methods: lessons from recent research* (pp. 184–200). Routledge.

Collins, D. (2014). *Cognitive interviewing practice.* Sage.

Conrad, F. G., Couper, M. P., & Sakshaug, J. W. (2016). Classifying open-ended reports: factors affecting the reliability of occupation codes. *Journal of Official Statistics*, *32*(1), 75–92.

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

Davidov, E., Dülmer, H., Cieciuch, J., Kuntz, A., Seddig, D., & Schmidt, P. (2016). Explaining measurement nonequivalence using multilevel structural equation modeling: the case of attitudes toward citizenship rights. *Sociological Methods & Research*, *Online first*. doi:10.1177/0049124116672678

Davidov, E., Meuleman, B., Cieciuch, J., Schmidt, P., & Billiet, J. (2014). Measurement equivalence in cross-national research. *Annual Review of Sociology*, *40*, 55–75.

Denscombe, M. (2008). The length of responses to open-ended questions. A comparison of online and paper questionnaires in terms of a mode effect. *Social Science Computer Review*, *26*(3), 359–368.

Fitzgerald, R., Widdop, S., Gray, M., & Collins, D. (2009). *Testing for equivalence using cross-national cognitive interviewing*. Centre for Comparative Social Surveys Working Paper Series, No. 1.

Goerman, P. L. & Caspar, R. A. (2010). Managing the cognitive pretesting of multilingual survey instruments: a case study of pretesting of the US Census Bureau bilingual Spanish/English questionnaire. In J. A. Harkness, M. Braun, B. Edwards, T. P. Johnson, L. Lyberg, P. P. Mohler, ... T. W. Smith (Eds.), *Survey methods in multinational, multiregional, and multicultural contexts* (pp. 75–90). Wiley-Blackwell.

Harkness, J. (2008). Comparative survey research: goals and challenges. In E. De Leeuw, J. Hox, & D. Dillman (Eds.), *International handbook of survey methodology* (pp. 56–77). Taylor & Francis.

He, J. & van de Vijver, F. (2012). Bias and equivalence in cross-cultural research. *Online Readings in Psychology and Culture*, *2*(2), 1–18.

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.

Jak, S., Oort, F. J., & Dolan, C. (2013). A test for cluster bias: detecting violations of measurement invariance across clusters in multilevel data. *Structural Equation Modeling, 20*, 265–282.

Jak, S., Oort, F. J., & Dolan, C. (2014). Measurement bias in multilevel data. *Structural Equation Modeling*, *21*, 31–39.

Jöreskog, K. (1971). Simultaneous factor analysis in several populations. *Psychometrika*, *36*(4), 409–426.

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*(3), 305–322.

Meitinger, K. (2017). Necessary but insufficient: why measurement invariance tests need online probing as a complementary tool. *Public Opinion Quarterly*, *81*(2), 447–472.

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

Meuleman, B. & Schlüter, E. (2018). Explaining cross-national measurement inequivalence: a Bayesian multilevel CFA with random loadings. In E. Davidov, P. Schmidt, J. Billiet, & B. Meuleman (Eds.), *Cross-cultural analysis: methods and applications.* (2nd ed., pp. 363–390). Routledge.

Miller, A. L. & Lambert, A. D. (2014). Open-ended survey questions: item nonresponse nightmare or qualitative data dream? *Survey Practice*, *7*(5).

Miller, K., Mont, D., Maitland, A., Altman, B., & Madans, J. (2011). Results of a cross-national structured cognitive interviewing protocol to test measures of disability. *Quality & Quantity*, *45*, 801–815.

Muthén, B. & Asparouhov, T. (2012). Bayesian structural equation modeling: a more flexible representation of substantive theory. *Psychological Methods*, *17*(3), 313–335.

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

Schulz, S., Meitinger, K., Braun, M., & Behr, D. (forthcoming). Who's bad? Eine Analyse zur internationalen Vergleichbarkeit von Maßen krimineller Einstellungen mittels des Web-Probing Ansatzes. In *Tagungsband der 15. Tagung der Kriminologischen Gesellschaft*.

Smith, T. W. (2010). The globalization of survey research. In J. A. Harkness, M. Braun, B. Edwards, T. P. Johnson, L. Lyberg, P. P. Mohler, ... T. W. Smith (Eds.), *Survey methods in multinational, multiregional, and multicultural contexts* (pp. 475–484). Wiley-Blackwell.

Tourangeau, R., Rips, R., & Rasinski, K. (2003). *The psychology of survey response*. Cambridge University Press.

Van de Vijver, F. & Leung, K. (2011). Equivalence and bias: a review of concepts, models, data analytic procedures. In D. Matsumoto & F. van de Vijver (Eds.), *Cross-cultural research methods in psychology* (pp. 17–45). Cambridge University Press.

Van de Vijver, F. & Poortinga, Y. (1997). Towards an integrated analysis of bias in cross-cultural assessment.

European Journal of Psychological Assessment, *13*(1), 29–37.

Willis, G. B. (2005). *Cognitive interviewing: a tool for improving questionnaire design*. Sage Publications.

Willis, G. B. (2015). The practice of cross-cultural cognitive interviewing. *Public Opinion Quarterly*, *79*(S1), 359–395.

Appendix A
Terms and Definitons

Table A1
Definitions of High and Low Education in the Five Countries for Quota Assignment

High	Low
Germany Fachhochschulreife, Abschluss einer Fachoberschule, Allgemeine oder fachgebundene Hochschulreife/Abitur (Gymnasium bzw. EOS, auch EOS mit Lehre)	**Germany** Von der Schule abgegangen ohne Hauptschulabschluss (Volksschulabschluss), Hauptschulabschluss (Volksschulabschluss), Realschulabschluss (mittlere Reife), Polytechnische Oberschule der DDR mit Abschluss der 8. oder 9. Klasse, Polytechnische Oberschule der DDR mit Abschluss der 10. Klasse,
Great Britain 2 or more A-levels, S-levels, A2-level, Scottish Highers, Scottish SCE/SLC/SUPE at Higher Grade, Scottish Higher School Certificate, Certificate of Sixth Year Studies/Advanced Higher Grade, Welsh Advanced Baccalaureate, Northern Ireland Senior Certificate, International Baccalaureate	**Great Britain** Skills for Life (including Basic Skills, Key Skills, Entry Level Certificates), 1-4 GCSEs A*-C, GCSE Grades D-G, Short course GCSE, CSE Grades 2-5, GCS O-level Grades D-E or 7-9, Scottish (SCE) Ordinary Bands D-E, Scottish Standard Grades 4-7, Scottish School Leaving Certificate - no grade, Scottish Access 1-3, Scottish Intermediate 1 (below A grade), GNVQ or GSVG Foundation level, Foundation Welsh Baccalaureate, 5 or more GCSEs A*-C, CSE Grade 1, GCE O-level Grades A-C or 1-6 Scottish SCE Ordinary Bands A-C or Pass, Scottish Standard Grades 1-3 or Pass, School Certificate or Matriculation Scottish School eaving certificate Lower Grade, SUPE Ordinary, Scottish Intermediate 1 (A grade), Scottish Intermediate 2, Intermediate Welsh Baccalaureate, Northern Irish Junior Certificate, 1 A-level or equivalent, Vocational GCSE, SCOTVEC/SQA National certificate modules/National Courses, BTEC First Certificate, GNVQ Intermediate
United States Associate, Junior College Bachelor's degree Graduate School	**United States** Less than high school, High school
Spain Bachillerato Superior (LOGSE), Arquitectura Técnica, Ingeniería Técnica (escuelas técnicas de tres años), Diplomaturas (tres años completos de cualquier carrera no técnica), Arquitectura e Ingeniería Licenciatura	**Spain** Menos de 5 aũos de escolarización, Enseũanza Primaria de LOGSE, ESO, EGB Bachillerato Elemental, Formación Profesional de grado medio (antigua FP1), Formación Profesional de grado superior
Mexico Titulaciones oficiales de estudios de postgrado (doctorado, especialidades médicas), Educación técnica posterior a secundaria Preparatoria completa, Educación técnica posterior a Bachillerato Carrera universitaria completa, MaestrÃŋa o doctorado	**Mexico** Ninguno o aún en la escuela, Primaria completa, Educación técnica sin secundaria, Secundaria completa

Table A2
Translation of the Closed Item (ISSP Original Language Versions Adapted to the Web Mode)

ISSP Item	Scale labels
Germany	
Und wie wichtig ist es für Sie, dass Menschen, die wegen schwerer Verbrechen verurteilt wurden, ihre Bürgerrechte verlieren?	1: überhaupt nicht wichtig 2: sehr wichtig DK: kann ich nicht sagen
Great Britain	
And how important is it that people convicted of serious crimes lose their citizen rights?	1: not at all important 7: very important DK: can't choose
United States	
And how important is it that people convicted of serious crimes lose their citizen rights?	1: not at all important 7: very important DK: can't choose
Spain	
Y ¿ hasta qué punto considera importante que las personas condenadas por delitos graves pierdan sus derechos de ciudadanÃŋa?	1: nada importante 7: muy importante DK: no sabe
Spain	
Y ¿ qué tan importante es que las personas condenadas por delitos graves pierdan sus derechos ciudadanos?	1: nada importante 7: muy importante DK: no sabe

We used the original translations of the 2013 ISSP questionnaires but adapted them to the web mode by adding "und/and/y" (marked in italics).

Table A3
Probe Translations for the Five Countries.)

Country	Probe
Category Selection Probe	
Germany	Bitte begründen Sie, warum Sie sich für "sehr wichtig" entschieden haben.
Great Britain	Please explain why you selected "very important."
United States	Please explain why you selected "very important."
Spain	Por favor, explique por qué se ha decidido por "muy importante".
Mexico	Por favor, explique por qué se decidió por "muy importante".
Specific Probe	
Germany	An welche Bürgerrechte haben Sie haben Sie bei der Beantwortung der Frage gedacht?
Great Britain	What particular citizen rights did you have in mind when you were answering the question?
United States	What particular citizen rights did you have in mind when you were answering the question?
Spain	¿En qué derechos de ciudadanía ha pensado Ud. al contestar a la pregunta?
Mexico	¿En qué derechos ciudadanos pensó Ud. al contestar a la pregunta?
Comprehension Probe	
Germany	Was verstehen Sie unter "schwere Verbrechen"?
Great Britain	What do you consider to be a "serious crime?"
United States	What do you consider to be a "serious crime?"
Spain	¿Qué entiende Ud. por "delitos graves"?
Mexico	¿Qué entiende Ud. por "delitos graves"?

Appendix B
Tables

Table B1

Mismatching Probe Responses for the Category-Selection Probe by Experimental Group and Country in Percent and detailed results for the two-sided Fisher's exact Test

Probe:		Category-selection					
Mismatch:		Specific			Comprehension		
	N	Categ.[a] %	Compr.[b] %	p	Categ.[a] %	Compr.[b] %	p
Germany	268	0.00	0.00	-	0.78	0.00	0.48
GB	277	0.00	1.33	0.50	0.00	0.00	-
U.S.	280	0.00	4.26	0.03	0.00	0.00	-
Spain	252	1.55	0.00	0.50	0.78	0.81	1.00
Mexico	277	0.00	1.55	0.22	0.00	2.33	0.10
Total	1,354	0.30	1.47	0.04	0.30	0.59	0.69

[a] First probe: category-selection [b] First probe: comprehension

Table B2

Mismatching Probe Responses for the Specific Probe by Experimental Group and Country in Percent and detailed results for the two-sided Fisher's exact Test

Probe:		Specific					
Mismatch:		Category-Selection			Comprehension		
	N	Categ.[a] %	Compr.[b] %	p	Categ.[a] %	Compr.[b] %	p
Germany	268	5.43	4.32	0.78	0.78	1.44	1.00
GB	277	3.15	6.67	0.27	5.51	2.00	0.19
U.S.	280	5.04	4.96	1.00	5.04	4.26	0.78
Spain	252	3.88	4.07	1.00	3.88	4.88	0.77
Mexico	277	4.05	9.30	0.09	4.05	3.10	0.76
Total	1,354	4.32	5.87	0.22	3.87	3.08	0.46

[a] First probe: category-selection [b] First probe: comprehension

Table B3
Mismatching Probe Responses for the Comprehension Probe by Experimental Group and Country in Percent and detailed results for the two-sided Fisher's exact Test

Probe:		Comprehension					
Mismatch:		Category-selection			Specific		
		Categ.[a]	Compr.[b]		Categ.[a]	Compr.[b]	
	N	%	%	p	%	%	p
Germany	268	2.33	12.95	0.00	0.00	0.72	1.00
GB	277	2.36	17.33	0.00	0.00	0.67	1.00
U.S.	280	0.72	17.73	0.00	0.72	0.71	1.00
Spain	252	2.33	22.76	0.00	0.00	0.00	-
Mexico	277	3.38	32.56	0.00	0.00	0.00	-
Total	1,354	2.23	20.38	0.00	0.15	0.44	0.62

[a] First probe: category-selection [b] First probe: comprehension