Article

Social Science Computer Review
2021, Vol. 39(3) 434-455
© The Author(s) 2019
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/0894439319859848
journals.sagepub.com/home/ssc
SAGE

Using Apples and Oranges to Judge Quality? Selection of Appropriate Cross-National Indicators of Response Quality in Open-Ended Questions

Katharina Meitinger[1], Dorothée Behr[2], and Michael Braun[2]

Abstract

Methodological studies usually gauge response quality in narrative open-ended questions with the proportion of nonresponse, response length, response time, and number of themes mentioned by respondents. However, not all of these indicators may be comparable and appropriate for evaluating open-ended questions in a cross-national context. This study assesses the cross-national appropriateness of these indicators and their potential bias. For the analysis, we use data from two web surveys conducted in May 2014 with 2,685 respondents and in June 2014 with 2,689 respondents and compare responses from Germany, Great Britain, the United States, Mexico, and Spain. We assess open-ended responses for a variety of topics (e.g., national identity, gender attitudes, and citizenship) with these indicators and evaluate whether they arrive at similar or contradictory conclusions about response quality. We find that all indicators are potentially biased in a cross-national context due to linguistic and cultural reasons and that the bias differs in prevalence across topics. Therefore, we recommend using multiple indicators as well as items covering a range of topics when evaluating response quality in open-ended questions across countries.

Keywords

response quality, open-ended questions, comparability, cross-national, cross-cultural, web probing, indicators

Research Into Narrative Open-Ended Survey Questions

Narrative open-ended questions in surveys have seen a renewed interest thanks to the rise of web-administered surveys, the ease of their implementation in these web surveys, and advances in

[1] Utrecht University, Utrecht, the Netherlands
[2] GESIS—Leibniz Institute for the Social Sciences, Mannheim, Germany

Corresponding Author:
Katharina Meitinger, Utrecht University, Utrecht, the Netherlands.
Email: k.m.meitinger@uu.nl

automated text analysis (Poncheri, Lindberg, Thompson, & Surface, 2008; Schonlau & Couper, 2016). Open-ended survey questions help to explore themes and respondent's thoughts without restricting the respondents to select from a list of predetermined answer categories. Open-ended questions are versatile in that they can be part of a actual survey itself, a methodological study, or a pretesting study (e.g., Behr, Meitinger, Braun, & Kaczmirek, 2017). The freedom of respondents comes with a price, though, which is increased response burden and thus potentially reduced response quality (e.g., nonresponse, Barrios, Villarroya, Borrego, & Ollé, 2011, Denscombe, 2008). To find a balance between the strengths and limitations of open-ended questions, many efforts are directed toward optimizing their design and reducing the threats to response quality. Several methodological studies have addressed these issues particularly for web surveys. Some investigated the impact of survey stimuli on response quality, such as the size of answer boxes (Behr, Bandilla, Kaczmirek, & Braun, 2014; Christian & Dillman, 2004; Emde & Fuchs, 2012; Israel, 2010; Maloshonok & Terentev, 2016; Smyth, Dillman, Christian, & McBride, 2009), the number of answer boxes (Fuchs, 2009; Hofelich Mohr, Sell, & Lindsay, 2016; Keusch, 2014), the use of motivational sentences (Kaczmirek, Meitinger, & Behr, 2017; Oudejans & Christian, 2010; Smyth et al., 2009), clarification features (Metzler, Kunz, & Fuchs, 2015), examples (Tourangeau, Conrad, Couper, & Ye, 2014), the position of the open-ended question in the web survey (Miller & Lambert, 2014), and counters indicating the number of characters written (Emde & Fuchs, 2012). Other studies investigated the impact of respondents' characteristics such as age, gender, or education (Andrews, 2005; Barrios et al., 2011; Denscombe, 2008; Miller & Lambert, 2014; Smyth et al., 2009; Zuell, Menold, & Körber, 2015). To assess the response quality of open-ended questions in methodological studies such as listed above, but also in other research activities, researchers need to decide on indicators. The most common indicators of response quality in research on open-ended questions in web surveys are (1) nonresponse, (2) response length, (3) response time, and (4) number of themes mentioned.

Most of the listed studies were conducted in only one country, and therefore, they used national and typically monolingual samples for their analyses. What happens, however, if researchers plan to extent their sample to include multilingual and multicultural respondents as well as several countries? Can they use the same set of indicators to assess the quality of responses given? This question is not trivial given that cross-national and cross-cultural studies are enjoying a great popularity (Smith, 2010), that these studies often adopt and/or adapt methods from general survey research, and that mixed-methods approaches, that is, quantitative and qualitative research, are of great value for comparative studies. After all, open-ended survey questions allow grasping the complexity of culture in cross-national or cross-cultural data (Van de Vijver & Chasiotis, 2010) and thus help to embed answers into their contexts. Since questionnaire design features and their universal applicability are increasingly being questioned (Silber, Stark, Blom, & Krosnick, 2019), it should not come as a surprise that also response quality indicators become the object of methodological studies.

With this article, we aim to raise awareness of what it means to determine response quality of open-ended answers in a cross-national and cross-cultural setting. We aim to look into the four indicators of response quality that are typically used in general survey research and assess their applicability in comparative research: Do these indicators work in comparative contexts or do linguistic and cultural variations in response behavior introduce a country bias in cross-national research on open-ended questions, which would then call for a more considered and informed use of indicators in cross-national and cross-cultural contexts? Ultimately, this research can contribute to efforts in cross-national studies such as the European Social Survey to define the quality of cross-national surveys through data quality assessments and data quality reports (http://www.europeansocialsurvey.org/methodology/ess_methodology/data_quality.html).

Are Indicators of Response Quality Transferable to the Cross-National Context?

The question of the transferability of indicators of response quality is crucial since it is unclear whether findings from methodological studies mainly conducted in the United States and other Western countries are generalizable to cultural contexts that are distinct from the western hemisphere (see also Stark et al., 2018).

Previous research on response behavior to closed questions revealed that culture partially influences the prevalence of response styles such as acquiescence or extreme response styles (Harzing, 2006; Johnson, Kulesa, Cho, & Shavitt, 2005) and that culture and country-level characteristics explain more of the variance of response styles in cross-cultural studies than sociodemographic and personality variables (Van Vaerenbergh & Thomas, 2012).

These findings suggest that cultural variations have an impact on responses, and similar differences that are driven by culture and country-level characteristics may be found for open-ended questions and their indicators, too.

Response Length

The most common indicator of response quality is response length (in words). Several methodological studies on open-ended questions assume that longer responses are better responses (e.g., Andrews, 2005; Denscombe, 2008). However, does this assumption also hold true for the cross-national context?

Differences in response length due to linguistic reasons. The difference in response length across countries is partially driven by the fact that languages, due to inherent grammatical features, differ in the number of words that are necessary to express the same opinion (Nettle, 2012). When looking at original texts and their translations, these differences become obvious and translate into what we call text expansion: For instance, a study by Wells et al. (2010) reported that the Spanish-language versions of their questionnaires in the United States were on average 15% longer than the English source questionnaires. Grisay (2002) reported for the Programme for International Student Assessment (PISA) 2000 field trial significant differences in length of the stimuli between the English and French assessment instruments.

The indicator of information density (InDe) captures grammatical information that languages need to provide (e.g., gender, pronouns, see Dryer & Haspelmath, 2013). InDe refers to the linguistic information per syllable and runs from 0 to 1 with 1 indicating the maximal InDe (Pellegrino, Coupé, & Marsico, 2011). InDe differs across languages. English is a language with a high InDe (.91), German has moderate InDe (.79), and Spanish has comparatively low InDe (.63). As a consequence, English-speaking respondents most likely need fewer characters to express the same information than Spanish-speaking respondents. Based on these theoretical considerations, we can formulate our first hypothesis.

> **Hypothesis 1:** Differences in InDe lead to shorter responses in English and longer responses in Spanish.

Additionally, InDe differences might translate into variations in the length of the (probe) question itself, which potentially affects response time (see further below).

Differences in response length due to cultural factors. Besides linguistic features, cultural factors such as culture-specific communication styles may drive cross-cultural differences in response length. Hall (1976) distinguishes between high- and low-context cultures with regard to the overtness of the message in communication. Countries vary in the amount of information that is "spelled out" and explicitly said in communication (Lim, 2002). In high-context countries, people tend to adopt an implicit communication style (Hall, 1976). Conversation partners tend to avoid straightforward messages, and conversation is characterized by reading between the lines and the use of metaphors. In low-context countries, the communication is more straightforward, and people often opt for an explicit communication style, that is simple, linear, and clear. However, there is to date no well-established, empirically founded country classification. Regardless, most previous studies classified the United States and Germany as low-context countries and Mexico as high-context country (Kittler, Rygl, & Mackinnon, 2011). Spain is located between the two extremes with tendencies toward high-context communication (Shao & Hill, 1994). For Great Britain, studies reported contradicting results, Rosenbloom and Larsen (2003) defined it as a low-context culture, but Djurssa (1994) identified elements of high-context communication in British business culture. Given that the communication style of low-context countries is often described as direct, accurate, precise, and explicit (Zaharna, 1995), cultural differences regarding the overtness of the message potentially translate into differences in response length.

A related concept to Hall's approach is the distinction between exacting and elaborate communication styles (Gudykunst, Ting-Toomey, & Chua, 1988).[1] This distinction addresses the quantity or volume of words that different cultures tend to use in communication (Neuliep, 2017). In countries where an elaborate communication style is dominant (e.g., Arabic and partly Latin American countries), conversations are characterized by a richness of language with many repetitions and an expressive and "flowery" language (Liu, 2016). In contrast, in countries with an exacting communication style (e.g., Germany), communication partners aim to provide the exact amount of information that is necessary to get the message across. Phrases follow the principle of "neither more nor less."

> **Hypothesis 2:** Responses in low-context countries and countries with an exacting communication style are shorter than responses in high-context countries and countries with an elaborate communication style.

Response Time

The interpretation of response time is not straightforward. First, response time can be interpreted both as an indicator of item difficulty and of increased response quality since respondents took their time to answer a question (Olson & Parkhurst, 2013). Although most of the research about response time is on closed questions, the same interpretation problem applies to open-ended questions. Second, linguistic reasons and cultural differences in time perception can potentially have an impact on response time.

Differences in response time due to linguistic reasons. As already discussed regarding response length, aspects such as text expansion and differential InDE lead to variations of text length in (translated) questions *and* written responses across languages. As a consequence, Spanish respondents most likely need to read longer questions *and* need to write longer texts, which will increase the reading and writing time (WT) necessary to complete an open-ended question.

> **Hypothesis 3:** Longer questions and responses translate into longer WT and response latencies.

Differences in response time due to cultural factors. Differences in time perception might create differences in response time that are partly driven by cultural factors. For closed items, Johnson, Holbrook, and Stavrakantonaki (2015) detected small differences in response latencies in the United States between White Americans, Mexican Americans, African Americans, and Korean Americans. They concluded that this might be due to cultural differences in time perceptions and time utilization. On the survey level, Wells, Vidalon, and DiSogra (2010) reported for the United States that the overall completion time of their survey was on average 75–150% longer in the Spanish compared to the English-language version. Although Spanish has a lower InDE than English, the expected text expansion due to translation is only 15% and, thus, inadequately accounts for this pronounced difference between both language versions. Wells et al. concluded that, in addition to the Spanish language, differences in culture and survey taking behaviors most likely increased the response length.

Nonresponse

Nonresponse to open-ended questions differs from nonresponse in closed questions: It can be either direct (hard) or indirect (soft). A hard nonresponse (HNR) would be a blank answer box (no written text at all). However, a respondent can also provide a soft nonresponse (SNR) by writing text that is of no use. SNR can be unintelligible letter combinations (e.g., "xcvbnm"), explicit refusals (e.g., "no comment"), don't knows, and meaningless or incomprehensible answers (e.g., "just cause"; Behr et al., 2014; Holland & Christian, 2009).

Rules of conduct in communication can differ across countries and potentially impact on nonresponse behavior. Cultures could have different expectations on how respondents should behave in an interview situation, for example, to which degree respondents should cooperate and how acceptable it is to reject a request directly. Such behavioral expectations may also appear in interview situations without interviewers, for example, in web surveys.

Differences in nonresponse due to cultural factors. In this context, one of Hofstede's dimensions of cultural variability may provide further insights:[2] the distinction between individualistic and collectivistic cultures (Gudykunst & Lee, 2003). Whereas in individualistic cultures the individual is more important than the group, in collectivistic cultures, the group takes precedence over the individual (Triandis, 1988). Respondents from collectivistic cultures try to avoid hurting others' feelings and do not impose their views or requests on others. Respondents from individualistic cultures, however, aim for clarity in conversations (Kim & Wilson, 1994). Cultures also differ as to how they react to requests: Individualistic cultures prefer to refuse directly, but collectivistic cultures are more indirect in their refusal and thus spare the person requesting loosing his or her face. Although Western cultures are usually classified as direct, there are variations regarding the level of Western directness (Lim, 2002; Wierzbicka, 1991), with Germans being particularly direct (House & Kasper, 1981). Although no interviewer is involved in a web survey, we suspect that communication styles partly drive response behavior in web surveys as well. Individualistic countries that are classified as direct may provide more nonresponses than collectivistic countries that are classified as more indirect.

> **Hypothesis 4**: Nonresponse is higher in individualistic countries than in collectivistic countries.

Also, country-specific rules of communication may have an impact on nonresponse behavior. For example, Triandis, Marín, Lisansky, and Betancourt (1984) observed that Hispanics follow the principle of simpatía during conversations. Conversation partners that obey this principle aim to be polite, likable, and respectful. Previous research showed that simpatía leads to increased levels of acquiescence among Hispanics for closed questions (Marin, Gamba, & Marin, 1992). However, simpatía may also reduce nonresponse at closed and open-ended questions. We argue that Mexican respondents potentially are "too polite" to refuse a response to an open-ended question, which could have an impact on nonresponse behavior.[3]

> **Hypothesis 5:** Nonresponse is lower in countries where the principle of simpatía is prevalent.

Number of Themes

The fourth indicator is the number of themes mentioned. A precondition for mentioning multiple themes is that respondents are familiar with the topic at hand. This might not always be the case since some concepts are emic (culture-specific) rather than etic (universal; Berry, 1969). If respondents in Country A are familiar with a concept but respondents in Country B are not, respondents in Country A probably come up with more associations regarding the concept. Additionally, the concepts' complexity might differ across countries. Concepts can entail a few aspects in one country but turn out to be multifaceted in another country.

Finally, questionnaire translation may also introduce bias in the number of themes, in particular if a translation leads to variation in the lexical scope of key terms in the question. For example, Meitinger (2017) identified issues of cross-national comparability in the translation of "social security benefits" in the International Social Survey Programme (ISSP) module on national identity, with the German translation having a large lexical scope ("sozialstaatliche Leistungen") encompassing any social security benefits and the English and Spanish version of this question triggering a more limited range of benefits.

> **Hypothesis 6:** Culture specificity of concepts has an impact on number of themes mentioned.

It is important to note that the same factors that might have a biasing impact on number of themes (applicability of concept, varying degree of complexity of concept, and translation) potentially affect the indicator of nonresponse as well. For example, a respondent not so familiar with a concept might also opt for a nonresponse instead of mentioning fewer topics.

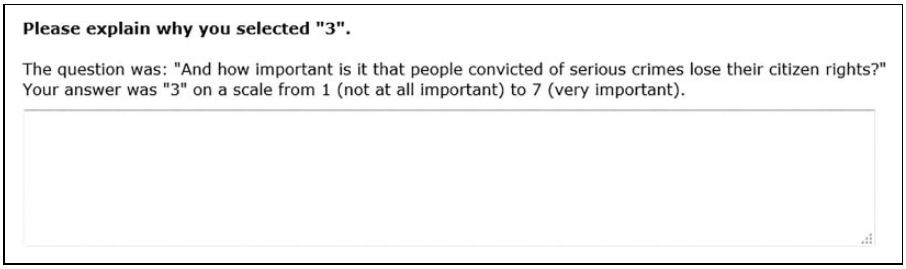
In this study, we aim to assess the cross-national appropriateness and potential bias of typical response quality indicators of open-ended questions. We are interested in whether we find cross-national differences and whether linguistically and culturally induced biases drive these differences. The latter would mean that the indicators will have to be used in a more considerate way when assessing response quality in a cross-national context. Furthermore, we are interested in whether the potential differences and biasing effects hold across topics and whether the indicators arrive at similar or contradictory conclusions regarding response quality.

Types of Open-Ended Questions

A variety of open-ended questions exist. Couper, Kennedy, Conrad, and Tourangeau (2011) distinguish between formatted numeric or verbal responses (e.g., questions asking for telephone numbers), frequency or numeric responses (e.g., "During the past 12 months, how many times have you seen or talked to a doctor about your health?"), single-word/phrase verbal responses (e.g., "What is your country of origin?"), short verbal responses with constraints on length but not format (e.g., "What kind of work was this person doing?"), and narrative responses with no length or formatting constraints (e.g., "What is the biggest problem facing the country today?").

Web Probes as a Particular Type of Open-Ended Question

In this study, we present results for a particular type of open-ended questions, namely, cognitive probes. They require narrative answers from the respondents but always in relation to a preceding



Please explain why you selected "3".

The question was: "And how important is it that people convicted of serious crimes lose their citizen rights?"
Your answer was "3" on a scale from 1 (not at all important) to 7 (very important).

Figure 1. Screenshot of a category-selection probe.

closed-ended question (Behr et al., 2017). Within Couper et al.'s (2011) typology, probes most closely resemble narrative responses. Probes are usually used at the pretest stage during cognitive interviewing at which survey questions are administered to respondents (Beatty & Willis, 2007). Web probing is the application of probing techniques from cognitive interviewing in web surveys to reveal respondents' cognitive processes when answering a survey question and—in a cross-national context—to uncover equivalence problems (Behr et al., 2017). Different types of probes can be distinguished: Category-selection probes ask about the reasons for having chosen an answer category, specific probes ask for additional information on a detail in the question, and comprehension probes request a definition of a key term (Prüfer & Rexroth, 2005; Willis, 2004). Figure 1 shows an example of a category-selection probe. The different probe types translate into slightly different responses. For instance, category-selection probes require that the respondents explain their perspective—which most likely leads to longer responses. In contrast, responses to specific probes are shorter because respondents often provide list-style responses (e.g., when asked which group of immigrants they had in mind, respondents provided answers such as "Turks, Italians, Chinese"; Behr et al., 2017). Therefore, we additionally indicate the probe type in our analysis. However, the most critical question in this context is whether there is a systematic response pattern across countries in addition to the variation due to probe type.

Data and Method

Data

For this study, we used data from two web surveys. Web Survey 1 was conducted in May 2014 with 2,685 respondents and replicated items from the ISSP Module on "national identity." Web Survey 2 was implemented in June 2014 with 2,689 respondents and replicated items from the ISSP Modules on "Citizenship" (ISSP Research Group, 2016a) and "family and gender roles" (ISSP Research Group, 2016b).[4] The country sample was Germany, Great Britain, Mexico, Spain, and the United States. Respondents came from nonprobability panels with quotas for age (18–30, 31–50, and 51–65), gender, and education (lower and higher). The main panel provider was respondi (www.respondi.com). The panel provider cooperated with international partners the choice of which we could not control. We met quotas in both surveys, Tables OS1 and OS2 in the Online Appendix provide detailed quota information. To ensure that the visual design for all respondents in the five countries is identical, we programmed the survey ourselves. The web surveys contained several experimental splits. Therefore, not all probes were answered by all respondents (see Table 1; for sample sizes per probe by country see Table OS3 in the Online Appendix). For the overall analysis (all probes collapsed), we created an aggregated data set in long format. As a consequence, the sample size in the overall analysis is based on the number of respondents times the received probes

Table 1. Variable Name, ISSP Item, Probe Type, Probe Wording, Number of Distinct Themes, Survey, and Sample Size for Each Open-Ended Probe.

Variable	ISSP Item	Probe Type	Probe	Distinct Themes	Survey	*N*
Gender	Consider a family with a child under school age. What, in your opinion, is the best way for them to organize their family and work life?	CSP	Please explain why you selected "Each family should find the solution which works best for them."	13	2	766
Pride	How proud are you of being British?	CSP	Please explain why you selected "very proud."	24	1	2,685
Democracy	How proud are you of Britain with regard to the way democracy works?	CSP	Please explain why you selected "very proud."	15	1	543
Social security	And how proud are you of Britain with regard to its social security system?	SP	What particular social security benefits did you have in mind when you were answering the question?	15	1	543
Treatment	And how proud are you of Britain with regard to its fair and equal treatment of all groups in society?	SP	What particular groups in society did you have in mind?	14	1	545
Patriotic feelings	How much do you agree or disagree that strong patriotic feelings in Britain strengthen Britain's place in the world?	COP	What do you associate with the phrase "strong patriotic feelings"?	36	1	2,685

Note. CSP = category-selection probe; SP = specific probe; COP = comprehension probe.

per respondents ($N = 7{,}767$). Additionally, we ran question-by-question analysis in the original data sets of Web Survey 1 and Web Survey 2.

Open-Ended Questions

We included six different probes in our analyses that cover the topics of family and gender roles and national identity. In our web surveys, we replicated the original ISSP items and their official ISSP translations. We developed and translated the probes ourselves following the Translation, Review, Adjudication, Pre-testing and Documentation (TRAPD) approach (Harkness, 2003). Table OS4 contains the exact question wording and translations for all probes. We also calculated the average number of characters of the probes for the different language versions with Britons and Americans receiving the shortest probes (128 characters) followed by Mexicans (148 characters), Germans (150 characters), and Spaniards (163 characters). We find variations in probe length for the questions for Spaniards and Mexicans (see also Table OS4).

For all probes, we developed a coding schema based on the probe responses (see also Meitinger, 2017, 2018; Meitinger, Braun, & Behr, 2018). The codes represent the different substantive themes mentioned and are the basis for the calculation of the indicator "number of themes mentioned" (see Table 1 for the maximum number of distinct themes for each coding schema). Since we developed the codes for each coding schema separately, the maximum number of codes that could be assigned varied across probes. All probes were coded and (partly) coded a second time to assess intercoder reliability, which was high for all probes (Holsti's coefficient: .86 to .98; calculated based on type and number of codes assigned).

Indicators & Analysis

In our study, we analyzed how the most frequently used indicators of response quality of open-ended questions behave in a cross-national setting. For each indicator, we report the overall values (all probes combined) and for each probe separately as well as their respective test statistic. (1) Our first indicator was *nonresponse.* HNR referred to respondents who left answer boxes blank. Most of the studies reporting nonresponse in open-ended questions refer to HNR (e.g., Miller & Lambert, 2014). SNR meant that the respondents provided a nonresponse even though they wrote some text. HNRs were easily detectable (no text entry). SNR was more difficult to analyze since it only became visible through coding. We coded all probe answers as SNR that had unintelligible letter combinations (e.g., “xcvbnm”), explicit refusals (e.g., “no comment”), don’t knows, and meaningless or incomprehensible answers (e.g., “just cause”; see Behr et al., 2014; Holland & Christian, 2009). We assessed whether there was a significant association between country and nonresponse (soft/hard) by running Pearson’s χ^2 tests. Cramer’s V serves as measure of effect size for this indicator. In total, we ran seven Pearson’s χ^2 tests (overall: 1, probes: 6). (2) Our second indicator is *response length*. To achieve a finer granulated comparison across languages, we assessed the average response length in number of characters instead of number of words. We controlled for outliers by replacing any values above the 95th percentile with the value of the 95th percentile. For each probe separately, we report median, mean, and analysis of variance (ANOVA) results for respondents who gave a substantive response (no nonresponse) including pairwise comparisons between countries using the Bonferroni correction. In addition, we report overall results of response length and response length adjusted for InDe on aggregated data (all probes combined). For the overall data, we conducted a two-way ANOVA (Factor 1: country, Factor 2: probe, interaction). (3) Our third indicator was *response time*. *Response latency* (RL) is the response time from screen load till first key stroke (reading and reflection time before starting to type the response). *Writing time* (WT) is the time between the first key stroke till pressing of the “continue” button (time of typing the response). We controlled for outliers by replacing unreasonably high values (above 5 min for RL and 10 min WT) and responses above the lowest and highest percentile with the value of these percentiles. Due to the nonnormal distribution of response time data, we report the median and the square root of the mean. We report ANOVA results for each probe and a two-way ANOVA (Factor 1: country, Factor 2: probe, interaction) for the overall data (all probes combined) including pairwise comparisons between countries using the Bonferroni correction for the square root transformed data. (4) Finally, we gauged response quality with the *number of themes mentioned.* Here, we compare the mean number of themes mentioned per respondent for each probe across countries and report the ANOVA results as well as the two-way ANOVA results for the analysis for all probes combined (overall).

Country Selection

Germany, Great Britain, the United States, Spain, and Mexico were part of this study. We covered the languages English, German, and Spanish with Spanish having the lowest and English having the highest InDE. From a cultural perspective, the countries differ regarding individualism/collectivism (United States: most individualistic; Mexico: most collectivistic). Regarding Hall’s dimension, the United States and Germany are low-context countries, and Mexico is a high-context country. Spain takes up an intermediate position, and Great Britain can be classified as either as low-context or a high-context culture (see Table 2).

Table 2. Characteristics of Countries.

Characteristic	Germany	Great Britain	United States	Spain	Mexico
Hofstede: Individualism–Collectivism	Individualistic	Highly individualistic	Highly individualistic	Slightly collectivistic	Collectivistic
Hall: low–high context	Lower	Lower/Higher	Lower	Higher	Higher
Language	German	English	English	Spanish	Spanish
Information density	.79	.91	.91	.63	.63

Note. Table based on Hofstede (https://geerthofstede.com/research-and-vsm/dimension-data-matrix/); Kittler, Rygl, and Mackinnon (2011); Pellegrino, Coupé, and Marsico (2011).

Table 3. Percentage of Hard Nonresponse (HNR) by Country and Probe.

	Germany		Great Britain		United States		Mexico		Spain		Total	
Probe	*n*	%	*n*	%	*n*	%	*n*	%	*n*	%	*n*	%
Overall	34	2.1	15	1.0	19	1.2	5	.3	11	.7	84	1.1
			$\chi^2(4, N = 7{,}767) = 28.20, p < .001$, Cramer's $V = .06$									
Gender	0	0	0	0	2	1.2	0	0	1	.6	3	.4
			$\chi^2(4, N = 766) = 4.57, p = .334$, Cramer's $V = .08$									
Pride	10	1.8	5	.9	8	1.5	0	0	3	.6	26	1.0
			$\chi^2(4, N = 2{,}685) = 11.80, p = .019$, Cramer's $V = .07$									
Democracy	2	1.9	1	.9	1	1.0	0	0	0	0	4	.7
			$\chi^2(4, N = 543) = 3.89, p = .421$, Cramer's $V = .08$									
Social security	2	1.9	3	2.7	3	3.0	2	1.7	1	.9	11	2.0
			$\chi^2(4, N = 543) = 1.53, p = .821$, Cramer's $V = .05$									
Treatment	3	2.8	1	.9	0	0	1	1.0	1	.9	6	1.1
			$\chi^2(4, N = 545) = 4.14, p = .388$, Cramer's $V = .09$									
Patriotic	17	3.1	5	.9	5	.9	2	.4	5	.9	34	1.3
			$\chi^2(4, N = 2{,}685) = 19.22, p = .001$, Cramer's $V = .08$									

Results

Nonresponse. Our first indicators were HNR and SNR. Respondents rarely provided HNR, regardless of country (see Table 3). Germans provided on average 2.1% HNR. Mexico and Spain showed even lower levels of HNR, with .3% and .7%, respectively. For some probes, respondents in some countries did not provide HNR at all (Germans, Britons: "gender"; Americans: "treatment"; Spaniards: "democracy"). Mexicans were particularly motivated since the probes for "gender," "pride," and "democracy" had no HNR. Apart from pride and "patriotic feeling" (pride: $p = .019$; patriotic feeling: $p = .001$), there was no significant association between country and HNR. However, these questions also had the largest sample sizes ($N = 2{,}689$), which might drive statistical significance. This is also reflected in the small effect sizes for all probes (range of Cramer's V: .05–.09). The low level of HNR was surprising. Panel policies probably encouraged respondents to provide some written text to receive the panel incentive because previous studies reported higher levels of HNR (e.g., Miller & Lambert, 2014: 32–76%).

In contrast, we found high levels of SNR (see Table 4), and levels varied by topic and country. SNR was low at the gender probe but higher at the remaining probes. Mexicans provided the least SNR. Also, Spaniards showed a comparatively low level of SNR. Germany, Great Britain, and the United States varied by topic in SNR (above 10% of respondents in Germany: "treatment," patriotic

Table 4. Percentage of Soft Nonresponse (SNR) by Country and Probe.

	Germany		Great Britain		United States		Mexico		Spain		Total	
Probe	*n*	%	*n*	%	*n*	%	*n*	%	*n*	%	*N*	%
Overall	137	8.8	157	10.0	173	11.23	36	2.4	83	5.3	586	7.5
			$\chi^2(4, N = 7{,}767) = 116.94, p < .001$, Cramer's $V = .12$									
Gender	5	3.5	4	2.4	1	.6	0	0	1	.6	11	1.4
			$\chi^2(4, N = 766) = 8.83, p = .066$, Cramer's $V = .11$									
Pride	52	9.4	51	9.5	57	10.7	10	1.9	39	7.3	209	7.8
			$\chi^2(4, N = 2{,}685) = 36.82, p < .001$, Cramer's $V = .12$									
Democracy	8	7.8	10	9.1	5	5.0	2	1.7	2	1.8	27	5.0
			$\chi^2(4, N = 543) = 10.74, p = .030$, Cramer's $V = .14$									
Social security	3	2.9	15	13.6	8	8.0	3	2.5	4	3.6	33	6.1
			$\chi^2(4, N = 543) = 17.29, p = .002$, Cramer's $V = .18$									
Treatment	12	11.1	10	8.8	16	14.4	3	3.1	6	5.3	47	8.6
			$\chi^2(4, N = 545) = 11.06, p = .026$, Cramer's $V = .14$									
Patriotic	57	10.3	67	12.5	86	16.2	18	3.4	31	5.8	259	9.7
			$\chi^2(4, N = 2{,}685) = 64.50, p < .001$, Cramer's $V = .16$									

feeling; Great Britain: "social security," patriotic feeling; the United States: pride, treatment, patriotic feeling). However, SNR was low for some questions in these countries (e.g., Germany: 2.9% social security). Depending on the topic, cross-national differences regarding this indicator might be more or less visible.

Despite the variations due to question topic, we found for all questions (except gender) a significant association between country and SNR. SNR was elevated in individualistic countries (Great Britain, the United States, to a certain degree Germany). In more collectivistic countries, we found lower levels of SNR (Mexico: lowest level).

This might be an indication of the influence of simpatía that encourages Mexicans to be polite, likable, and respectful in conversations. Mexicans may be "too polite" to refuse a response.

The difference between the prevalence of HNR and SNR is of importance for methodological cross-cultural studies on open-ended question. Most methodological studies opt for an evaluation of response quality with only HNR; thus, the percentage of overall nonresponse could be underestimated. The cross-national differences regarding nonresponse behavior may be overlooked.

Response Length. Regarding response length, we found (see Table 5) that the average number of characters varied by probe type. Responses to category-selection probes (gender, pride, democracy) were longer than responses to specific (social security, treatment) and comprehension (patriotic feelings) probes. Also, length varied by topic. For example, the responses to the gender category-selection probe were longer than for pride and democracy category-selection probes.

Overall, we also find significant but small differences between the countries, $F(4, 7387) = 25.94$, $p < .001$, $\eta^2 = .01$. U.S. respondents provided the shortest responses and pairwise comparisons with each of the other countries indicated significantly different values. British respondents and German respondents wrote answers of medium length. Overall, Spanish and Mexican responses were the longest. Interestingly, country differences depended on topic, $F(20, 7387) = 2.20$, $p = .002$, $\eta^2 = .01$. For example, we did not find any significant differences for social security. Country differences for the other questions were all significant and showed small effect sizes ($\eta^2 = .02$ to $.04$).

The results reflect InDE differences of the respondents' languages. English-language responses turned out to be the shortest, which was expected given the high InDE level of English. Due to the

Table 5. Median and Mean Response Length (Standard Deviation) for Each Probe by Country for Substantive Respondents.

			Germany	Great Britain	United States	Mexico	Spain	Total	*F*(*df*)	*p*	η^2
Overall	Type	Median	49	47	41	77	70	55	25.94		
		M (*SD*)	80 (88)^{ab}	74 (76)^{acd}	65 (71)^{bef}	98 (78)^{ce}	96 (90)^{df}	83 (82)	(4, 7387)	<.001	.01
Gender	CSP	Median	120	98	90	119	128.5	109	3.13		
		M (*SD*)	149 (105)	130 (100)	119 (93)	150 (98)	147 (94)	138 (98)	(4, 750)	.015	.02
Pride	CSP	Median	74	63	54	93	92	75	23.49		
		M (*SD*)	97 (78)^{abc}	91 (79)^{def}	76 (71)^{adgh}	116 (80)^{beg}	117 (85)^{cfh}	100 (80)	(4, 2451)	<.001	.04
Democracy	CSP	Median	70	63	56	92	74	72.5	4.44		
		M (*SD*)	108 (91)^{a}	83 (60)	76 (67)^{abc}	106 (74)^{b}	111 (89)^{c}	98 (78)	(4, 507)	.002	.03
Social security	SP	Median	49	54.5	57	62	77	59	1.06		
		M (*SD*)	81 (88)	97 (97)	82 (77)	95 (87)	99 (80)	91 (86)	(4, 494)	.375	.01
Treatment	SP	Median	42	27	20	47.5	44	38	4.88		
		M (*SD*)	58 (52)	50 (55)	37 (42)^{ab}	66 (55)^{a}	62 (52)^{b}	55 (52)	(4, 487)	.001	.04
Patriotic feeling	COP	Median	31	36	33	54	41	39	17.61		
		M (*SD*)	49 (49)^{ab}	51 (46)^{cd}	49 (47)^{ef}	70 (52)^{ace}	62 (56)^{bdf}	57 (51)	(4, 2388)	<.001	.03
Overall adjusted for InDe		Median	39	43	37	49	44	42	5.55		
		M (*SD*)	63 (70) ^{ab}	67 (70)^{cd}	60 (64)	62 (49)^{ac}	60^{bd} (49)	60 (57)	(4, 7387)	<.001	.001

Note. Overall and overall adjusted for InDe: Two-way ANOVA (country and probe topic); gender–patriotic feeling: ANOVA; Pairwise comparisons between the countries using the Bonferroni correction. Significantly different (<.05) pairs of values are indicated by matching superscript letters. InDe = indicator of information density.

low InDE of Spanish, responses from Spain and Mexico were the longest. The medium InDE level of German also led to responses of intermediary length. To assess the impact of InDe on response length, we reran the analysis with response length adjusted for InDe. Although there is still a significant country difference, $F(4, 7387) = 5.55$, $p < .001$, effect sizes dropped ($\eta^2 = .001$).

The low-context countries with an exacting communication style provided shorter responses than high-context countries with an elaborate communication style.

Response Time

Our third indicator of response quality was response time in seconds. We distinguished between RL and WT.

RL. When comparing the time substantive respondents took from screen load till first key stroke (see Table 6), we found again overall small but significant cross-national differences, $F(4, 6{,}971) = 15.52$, $p < .001$, $\eta^2 = .01$. Overall, British respondents took the shortest amount of time for reading and thinking about their answers before starting to write their response (significant pairwise comparison with Mexico, and Spain). German and American respondents also had a short RL, whereas Spanish and especially Mexican respondents showed slightly longer RLs. The short RL of the English-speaking respondents is in line with the shorter English question text. Although Mexican respondents had the longest RL, their average question text was shorter than the German and the Spanish version. In addition, the country effect on RL also varied by question topic, $F(20, 6971) = 3.67$, $p < .001$, $\eta^2 = .01$.

WT. WT, that is, the time between the first key stroke and the click on the next button, varied by probe type and topic (see Table 7). Respondents in all countries had longer WTs at category-selection and specific probes (total square root transformed mean value [*SQM*]: .20 to.23) than at the comprehension probe (total *SQM*: .17). The question topic also influenced overall WT (longest total *SQM*: gender: .23; shortest: patriotic feeling: .17). Unfortunately, it was impossible to disentangle the influence of topic from probe type.

Overall, there is a small but significant country effect for all questions combined, $F(4, 6971) = 25.57$, $p < .001$, $\eta^2 = .02$. Americans had shortest WTs and Mexicans the longest WT, followed by Spaniards (all pairwise comparisons significant for Mexico and Spain). Britons and Germans took the middle position.

Except for social security, we found significant differences in WT for each question separately. The fact that the effect of social security is again not significant is intriguing. The item seems to behave differently than the other questions since we also did not find any significant effect for the indicators HNR and response length for this indicator.

Number of Themes

Overall (see Table 8), we found a small but significant country difference for number of theme, $F(4, 6971) = 9.77$, $p < .001$, $\eta^2 = .01$. U.S. respondents mentioned the fewest (pairwise comparisons significant for all countries), and Mexican respondents mentioned the most themes, followed by Spain. Germany and Great Britain took the middle position.

However, country differences varied along probes. The probes for gender and democracy triggered a similarly low number of themes in all countries, with no significant country differences for gender.

In contrast, social security revealed many themes in Germany, and in the pairwise comparisons, the value was significantly different from all other countries. A likely explanation is the larger

Table 6. Response Time–Response Latency: Median and Square Root Transformed Mean of Response Latency by Country in Seconds for Substantive Respondents Plus ANOVA Results.

Probe	Type		Germany	Great Britain	United States	Mexico	Spain	Total	*F*(*df*)	*p*	η^2
Overall		Median	6.80	6.27	6.55	8.89	7.27	7.13			
		SQM(*SD*)	.09 (.05)^{a}	.09 (.05)^{bc}	.10 (.05)^{d}	.11 (.06)^{abde}	.10 (.05)^{ce}	.09 (.05)	15.52 (4, 6971)	<.001	.01
Gender	CSP	Median	7.66	6.48	5.63	8.44	6.44	6.67			
		SQM (*SD*)	.10 (.04)	.09 (.04)	.09 (.05)	.10 (.04)	.09 (.05)	.09 (.04)	1.45 (4, 750)	.22	.01
Pride	CSP	Median	5.66	4.69	4.89	6.12	5.40	5.36			
		SQM (*SD*)	.09 (.04)	.08 (.04)^{a}	.08 (.05)	.09 (.05)^{a}	.09 (.04)	.08 (.04)	4.76 (4, 2,418)	.001	.01
Democracy	CSP	Median	7.33	4.94	5.21	7.02	7.10	6.39			
		SQM (*SD*)	.09 (.04)	.09 (.06)	.09 (.05)	.09 (.05)	.10 (.05)	.09 (.05)	.24 (4, 507)	.914	.00
Social security	SP	Median	7.12	8.32	9.50	11.63	7.77	8.59			
		SQM (*SD*)	.10 (.05)	.09 (.04)^{a}	.11 (.05)	.12 (.06)^{a}	.10 (.05)	.10 (.05)	2.94 (4, 381)	.021	.03
Treatment	SP	Median	9.56	8.64	8.01	14.23	13.29	9.99			
		SQM (*SD*)	.11 (.05)^{a}	.10 (.04)^{bc}	.10 (.05)^{de}	.13 (.07)^{abd}	.12 (.05)^{ce}	.11 (.05)	7.29 (4, 532)	<.001	.05
Patriotic feel	COP	Median	6.97	7.43	8.59	12.28	8.71	8.62			
		SQM (*SD*)	.10 (.05)^{ab}	.10 (.05)^{cd}	.11 (.06)^{ace}	.13 (.07)^{bdef}	.11 (.06)^{f}	.11 (.06)	26.30 (4, 2,383)	<.001	.04

Note. Overall: Two-way ANOVA (country and probe topic); gender–patriotic feeling: ANOVA; Pairwise comparisons between the countries using the Bonferroni correction. Significantly different (<.05) pairs of values are indicated by matching superscript letters. *SQM* = total square root transformed mean value; ANOVA = analysis of variance.

Table 7. Response Time–Writing Time: Median and Square Root Transformed Mean of Writing Time by Country in Seconds for Substantive Respondents Plus ANOVA Results

Probe	Type		Germany	Great Britain	United States	Mexico	Spain	Total	*F*(*df*)	*p*	η^2
Overall		Median	25.51	25.69	21.90	42.21	34.50	29.34			
		SQM (*SD*)	.18 (.10)^{abc}	.18 (.10)^{cde}	.17 (.10)^{afg}	.22 (.11)^{bdfh}	.20 (.11)^{cegh}	.20 (.12)	29.57 (4, 6971)	<.001	.02
Gender	CSP	Median	51.01	40.55	37.63	55.86	51.39	46.87			
		SQM (*SD*)	.24 (.12)^{a}	.24 (.13)	.21 (.09)^{abc}	.25 (.11)^{b}	.24 (.12)^{c}	.23 (.12)	3.80 (4, 750)	.005	.02
Pride	CSP	Median	30.67	30.32	22.62	49.31	38.71	33.43			
		SQM (*SD*)	.20 (.10)^{abc}	.20 (.11)^{def}	.18 (.11)^{adgh}	.24 (.12)^{beg}	.23 (.12)^{cfh}	.21 (.11)	29.75 (4, 2418)	<.001	.07
Democracy	CSP	Median	27.55	28.94	24.60	41.23	39.04	32.29			
		SQM (*SD*)	.22 (.13)	.19 (.09)	.19 (.11)	.23 (.11)	.22 (.11)	.21 (.11)	2.55 (4, 507)	.038	.02
Social security	SP	Median	24.57	28.59	28.82	40.70	35.80	32.73			
		SQM (*SD*)	.21 (.14)	.20 (.12)	.20 (.12)	.23 (.13)	.22 (.11)	.21 (.12)	.77 (4, 381)	.544	.01
Treatment	SP	Median	32.76	25.80	19.43	45.93	42.47	32.05			
		SQM (*SD*)	.20 (.12)^{a}	.18 (.09)^{b}	.17 (.11)^{cd}	.24 (.10)^{abc}	.22 (.10)^{d}	.20 (.10)	8.32 (4, 532)	<.001	.06
Patriotic feeling	COP	Median	14.04	17.98	16.09	30.69	21.56	19.71			
		SQM (*SD*)	.14 (.09)^{ab}	.16 (.10)^{c}	.15 (.09)^{de}	.20 (.11)^{acdf}	.17 (.11)^{bef}	.17 (.10)	26.40 (4, 2383)	<.001	.04

Note. Overall: Two-way ANOVA (country and probe topic); gender–patriotic feeling: ANOVA; Pairwise comparisons between the countries using the Bonferroni correction. Significantly different (<.05) pairs of values are indicated by matching superscript letters. CSP = category-selection probe; SP = specific probe; COP = comprehension probe; *SQM* = total square root transformed mean value; ANOVA = analysis of variance.

Table 8. Mean Number (Standard Deviation) of Mentioned Themes for Each Question by Country for Substantive Respondents.

Probe	Type	Germany	Great Britain	United States	Mexico	Spain	Total	*F*(*df*)	*p*	η^2
Overall		1.57 (.91)^{a}	1.62 (.97)^{b}	1.43 (.76) ^{abcd}	1.84 (1.12) ^{c}	1.72 (1.07)^{d}	1.64 (.99)	9.77 (4, 6971)	<.001	.01
Gender	CSP	1.26 (.58)	1.28 (.59)	1.18 (.53)	1.25 (.61)	1.29 (.58)	1.25 (.58)	.92 (4, 750)	.454	.01
Pride	CSP	1.81 (1.06)^{ab}	1.92 (1.11)^{cde}	1.62 (.80)^{cfg}	2.22 (1.26)^{adf}	2.17 (1.22)^{beg}	1.95 (1.25)	24.73 (4, 2418)	<.001	.04
Democracy	CSP	1.19 (.42)	1.34 (.54)^{a}	1.22 (.47)	1.15 (.36)^{a}	1.28 (.58)	1.24 (.48)	2.57 (4, 507)	.037	.02
Social security	SP	1.97 (1.15)^{abcd}	1.48 (.96)^{a}	1.17 (.41)^{b}	1.29 (.66)^{c}	1.29 (.67)^{d}	1.43 (.84)	11.95 (4, 381)	<.001	.11
Treatment	SP	1.85 (1.13)^{a}	1.81 (1.33)	1.42 (.79)^{a}	1.57 (.92)	1.77 (1.07)	1.69 (1.08)	3.17 (4, 532)	.014	.02
Patriotic feeling	COP	1.38 (.66)^{ab}	1.46 (.79)^{c}	1.40 (.81)^{de}	1.90 (1.13)^{acdf}	1.60 (1.01)^{bef}	1.56 (.92)	27.27 (4, 2383)	<.001	.04

Note. Overall: Two-way ANOVA (country and probe topic); gender-patriotic feeling: ANOVA; Pairwise comparisons between the countries using the Bonferroni correction. Significantly different (<.05) pairs of values are indicated by matching superscript letters. CSP = category-selection probe; SP = specific probe; COP = comprehension probe; ANOVA = analysis of variance.

lexical scope of this question as well as the broad range of services and benefits in Germany (Meitinger, 2017). The question is, therefore, a good example where translation and differences in the social security system may have a biasing impact on the indicator of number of themes. However, this probe also performed differently than the other probes regarding HNR, response length, and WT (no significant country difference). At the same time, Germany took a middle position with the other indicators, and frequently, pairwise comparison did not indicate significantly different values. It seems that the lexical scope is counteracting country differences in other dimensions of response quality (e.g., nonresponse, response length) for this item.

Mexicans mentioned most themes at the pride question (statistically significant pairwise comparisons with Germany, Great Britain, the United States). Given the substantive analysis of this question, this did not come as a surprise. Mexican respondents mentioned, in general, a larger variety of sources of national pride than respondents from other countries (Meitinger, 2018). Here, we have another intriguing example where the complexity of the concept varied across countries, which again potentially biases the indicator of number of themes.

Conclusion

We set out to explore whether we can find cross-national differences with regard to four indicators of response quality in open-ended questions and whether these variations hold across a range of different topics. More importantly, we wanted to evaluate whether some of these indicators are biased due to linguistic and cultural factors.

Overall, we find clear cross-cultural differences on the basis of the four indicators of response quality. Respondents from the United States and Mexico showed an opposite response behavior. Mexicans excelled at nearly all indicators of response quality (longest responses, low nonresponse, longest response time). Americans wrote the shortest responses, mentioned the fewest themes, responded fast, and also provided—for most topics—the highest nonresponse. Germany, Great Britain, and Spain were in between with Spain being located closer to Mexico and performing rather high on all indicators of response quality.

Do We Find Indications for Bias at the Different Indicators?

Although we revealed a striking cross-national response pattern, we also found indications that all indicators seem to be somewhat biased. The cross-national differences in response length parallel the linguistic differences between the countries regarding InDE. This explains why Spaniards and Mexicans wrote longer and Britons and Americans shorter responses. As we showed, country differences are reduced when InDE is accounted for in the analysis. The country pairs also differ regarding their communication style, with the United States being a low-context country, and Spaniards and Mexicans tending toward high-context communication. Also, we found indications that the number of themes may be a problematic cross-national indicator. Translations and differences in the social security system introduced some bias (varying lexical scope of social security) and this bias also spilled over to other dimensions of response quality and suppressed country differences for this probe. Differences in the complexity of concepts (sources of pride) biased this indicator, too. Finally, nonresponse was driven by the question topic: Depending on which topic is selected to evaluate response quality, the researcher might come to different conclusion. Furthermore, cultural values such as simpatía might reduce nonresponse in Mexico independent of the question topic. Therefore, we found clear indications that linguistic and cultural factors potentially drove responses—hence, results regarding response quality may be biased.

Does This Mean That We Should Completely Disregard All Indicators of Response Quality When Assessing Cross-National Data?

Although the indicators are potentially biased, they each shed light on different aspects of response quality. Instead of disregarding these indicators, we recommend using multiple indicators of response quality for the assessment of cross-national data and paying attention to linguistic and cultural factors in their interpretation. For instance, linguistic differences can be accounted for by introducing InDE as a control variable in statistical analyses of response length and time.

Also, we recommend assessing response quality in a methodological study with several questions covering different topics. Optimally, these questions are selected by survey and country experts who should pay attention that the topics of the questions do not differ across countries in terms of concept complexity and lexical scope of the translations. The probe of social security was a good example for a biased question where differences in lexical scope created a suppression effect for cross-national differences on other dimension of response quality. A careful selection of questions reduces the bias of the indicator of number of themes and nonresponse.

Limitations and Future Research

Several limitations of this study need to be acknowledged. First, this study analyzed a relatively small number of probes. This is due to the work-intensive data analysis of cross-national narrative open-ended questions. Second, due to data availability we compared English, German, and Spanish-speaking countries. The discussed issues may be even larger for non-Western languages, for example, Chinese, Arabic, or Japanese. Differences in responses styles are also likely to be larger when these countries are included. Third, with narrative open-ended probes, we tested one specific type of open-ended questions. Fourth, we analyzed open-ended probes in web surveys. We expect that cultural particularities might have a larger impact in face-to-face settings due to the presence of an interviewer. And finally, we cannot exclude effects specific to the use of online access panels. Therefore, future research should assess these effects with a larger number of questions covering a broader variety of topics and including different modes and countries outside the western hemisphere.

Data Availability

The data set generated and analyzed during the current study is available on request from the corresponding author. Email: k.m.meitinger@uu.nl

Software Information

All analyses in the present study were conducted using STATA Version 14.

Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: 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].

Supplemental Material

Supplemental material for this article is available online.

Notes

1. Gudykunst, Ting-Toomey, & Chua (1988) also discuss a succinct communication style. However, this communication style does not provide further insights for this article.
2. Hofstede's approach is not free of criticism. Schmitz and Weber (2014) found for the dimension of uncertainty avoidance a lack of configural invariance when conducting measurement invariance tests for a French and German sample. Therefore, we refrain from using the uncertainty avoidance dimension. However, Hofstede's dimensions have proven to be useful exploratory factors in cross-national methodological studies on topics such as social desirable responding (e.g., Middleton & Jones, 2000), acquiescent responding (e.g., Harzing, 2006), or extreme response style (De Jong, Steenkamp, Fox, & Baumgartner, 2008; Harzing, 2006; Johnson et al., 2005).
3. An alternative explanation for difference in nonresponse is survey fatigue (Goyder, 1986), which states that the more respondents have been asked to participate in surveys, the less likely they will take part in additional surveys or provide answers.
4. The data set generated and analyzed during the current study as well as the syntax are available on request from the corresponding author. Email: k.m.meitinger@uu.nl.

References

Andrews, M. (2005, May). *Who is being heard? Response bias in open-ended responses in a large government employee survey*. Paper presented at the 60th AAPOR Annual Conference, Miami Beach, FL.

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

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

Behr, D., Bandilla, W., Kaczmirek, L., & Braun, B. (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. *GESIS—Survey Guidelines*. Mannheim, Germany: GESIS—Leibniz-Institute for the Social Sciences.

Berry, J. W. (1969). On cross-cultural comparability. *International Journal of Psychology*, *4*, 119–128.

Christian, L. M., & Dillman, D. A. (2004). The influence of graphical and symbolic language manipulations on responses to self-administered questions. *Public Opinion Quarterly*, *68*, 57–80.

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

De Jong, M. G., Steenkamp, J.-B. E., Fox, J.-P., & Baumgartner, H. (2008). Using item response theory to measure extreme response style in marketing research: A global investigation. *Journal of Marketing Research*, *45*, 104–115.

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*, 359–368.

Djurssa, M. (1994). North European business cultures: Britain vs. Denmark and Germany. *European Management Journal*, *12*, 138–146.

Dryer, M. S., & Haspelmath, M. (2013). The World Atlas of language structures online. Max Planck institute for evolutionary anthropology, Leipzip. Retrieved February 5, 2018, from http://wals.info

Emde, M., & Fuchs, M. (2012). *Using adaptive questionnaire design in open-ended questions: A field experiment*. American Association for Public Opinion Research (AAPOR) 67th Annual Conference, San Diego, CA.

Fuchs, M. (2009). Differences in the visual design language of paper-and-pencil surveys versus web surveys: A field experimental study on the length of response fields in open-ended frequency questions. *Social Science Computer Review*, *27*, 213–227.

Goyder, J. (1986). Surveys on surveys: Limitations and potentialities. *Public Opinion Quarterly*, *50*, 27–41.

Grisay, A. (2002). Translation and cultural appropriateness of the test and survey material. In R. Wu (Ed.), *PISA* (pp. 57–70). Paris, France: Organization for Economic Cooperation and Development.

Gudykunst, W. B., & Lee, C. M. (2003). Cross-cultural communication theories. In W. B. Gudykunst (Ed.), *Cross-cultural and intercultural communication* (pp. 7–33). Thousand Oaks, CA: Sage.

Gudykunst, W. B., Ting-Toomey, S., & Chua, E. (1988). *Culture and interpersonal communication*. Thousand Oaks, CA: Sage.

Hall, E. T. (1976). *Beyond culture*. Garden City, NY: Anchor Press.

Harkness, J. A. (2003). Questionnaire translation. In J. A. Harkness, F. J. Van de Vijver, & P. Mohler (Eds.), *Cross-cultural survey methods* (pp. 35–56). New York, NY: John Wiley.

Harzing, A. W. (2006). Response styles in cross-national survey research: A 26-country study. *International Journal of Cross Cultural Management*, *6*, 243–266.

Hofelich Mohr, A., Sell, A., & Lindsay, T. (2016). Thinking inside the box: Visual design of the response box affects creative divergent thinking in an online survey. *Social Science Computer Review*, *34*, 347–359.

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*, 196–212.

House, J., & Kasper, G. (1981) Politeness markers in English and German. In F. Coulmas (Ed.), *Conversational routine: Explorations in standardized communication situations and pre-patterned speech* (pp. 157–185). The Hague, the Netherlands: Mouton.

Israel, G. D. (2010). Effects of answer space size on responses to open-ended questions in mail surveys. *Journal of Official Statistics*, *26*, 271–285.

ISSP Research Group. (2016a). International social survey programme Citizenship II—ISSP 2014 (ZA6670 Data file Version 2.0.0). GESIS Data Archive, Cologne. doi:10.4232/1.12590

ISSP Research Group. (2016b). International social survey programme family and changing gender roles IV—ISSP 2012 (ZA5900 Data file Version 4.0.0). GESIS Data Archive, Cologne. doi:10.4232/1.12661

Johnson, T., Holbrook, A., & Stavrakantonaki, M. (2015). *Cross-cultural equivalence of survey response latencies*. 6th Conference of the European Survey Research Association (ESRA). Reykjavik, Iceland.

Johnson, T., Kulesa, P., Cho, Y. I., & Shavitt, S. (2005). The relation between culture and response styles: Evidence from 19 countries. *Journal of Cross-Cultural Psychology*, *36*, 264–277.

Kaczmirek, L., Meitinger, K., & Behr, D. (2017). Higher data quality in web probing with EvalAnswer: A tool for identifying and reducing nonresponse in open-ended questions. GESIS—Leibniz-Institut für Sozialwissenschaften, Cologne. Retrieved from http://nbn-resolving.de/urn:nbn:de:0168-ssoar-51100-0

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.

Kim, M. S., & Wilson, S. R. (1994). A cross-cultural comparison of implicit theories of requesting. *Communications Monographs*, *61*, 210–235.

Kittler, M. G., Rygl, D., & Mackinnon, A. (2011). Special review article: Beyond culture or beyond control? Reviewing the use of Hall's high-/low-context concept. *International Journal of Cross Cultural Management*, *11*, 63–82.

Lim, T.-S. (2002). Language and verbal communication across cultures. In W. B. Gudykunst (Ed.), *Handbook of international and intercultural communication* (pp. 69–87). Thousand Oaks, CA: Sage.

Liu, M. (2016, November). Verbal communication styles and culture. *Communication and Culture Online*. doi: 10.1093/acrefore/9780190228613.013.162

Maloshonok, N., & Terentev, E. (2016). The impact of visual design and response formats on data quality in a web survey of MOOC students. *Computers in Human Behavior*, *62*, 506–515.

Marin, G., Gamba, R. J., & Marin, B. V. (1992). Extreme response style and acquiescence among Hispanics: The role of acculturation and education. *Journal of Cross-cultural Psychology*, *23*, 498–509.

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

Meitinger, K. (2018). What does the general national pride item measure? Insights from web probing. *International Journal of Comparative Sociology*, *59*, 428–450.

Meitinger, K., Braun, M., & Behr, D. (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.

Metzler, A., Kunz, T., & Fuchs, M. (2015). The use and positioning of clarification features in web surveys. *Psihologija*, *48*, 379–408.

Middleton, K. L., & Jones, J. L. (2000). Socially desirable response sets: The impact of country culture. *Psychology & Marketing Journal*, *17*, 149–163.

Miller, A. L., & Lambert, D. A. (2014). Open-ended survey questions: Item nonresponse nightmare or qualitative data dream? *Survey Practice*, *7*, 1–11.

Nettle, D. (2012). Social scale and structural complexity in human languages. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *367*, 1829–1836.

Neuliep, J. W. (2017). *Intercultural communication: A contextual approach*. Thousands Oaks, CA: Sage.

Olson, K., & Parkhurst, B. (2013). Collecting paradata for measurement error evaluations. In F. Kreuter, (Ed.), *Improving surveys with paradata: Analytic uses of process information* (pp. 43–72). Hoboken, NJ: John Wiley.

Oudejans, M., & Christian, L. M. (2010). Using interactive features to motivate and probe responses to open-ended questions. In M. Das, P. Ester, & L. Kaczmirek (Eds.), *Social and behavioral research and the Internet: Advances in applied methods and research strategies* (pp. 215–244). New York, NY: Routledge.

Pellegrino, F., Coupé, C., & Marsico, E. (2011). Across-language perspective on speech information rate. *Language*, *87*, 539–558.

Poncheri, R. M., Lindberg, J. T., Thompson, L. F., & Surface, E. A. (2008). A comment on employee surveys: Negativity bias in open-ended responses. *Organizational Research Methods*, *11*, 614–630.

Prüfer, P., & Rexroth, M. (2005). Kognitive Interviews [cognitive interviews]. *ZUMA How-to-Reihe, 15*. Retrieved January 10, 2018, from http://www.gesis.org/fileadmin/upload/forschung/publikationen/gesis_reihen/howto/How_to15PP_MR.pdf?download=true

Rosenbloom, B., & Larsen, T. (2003). Communication in international business-to-business marketing channels: Does culture matter? *Industrial Marketing Management*, *32*, 309–315.

Schmitz, L., & Weber, W. (2014). Are Hofstede's dimensions valid? A test for measurement invariance of uncertainty avoidance. *Interculture Journal: Online-Zeitschrift für interkulturelle Studien*, *13*, 11–26.

Schonlau, M., & Couper, M. P. (2016). Semi-automated categorization of open-ended questions. *Survey Research Methods*, *10*, 143–152.

Shao, A. T., & Hill, J. S. (1994). Global television advertising restrictions: The case of socially sensitive products. *International Journal of Advertising*, *13*, 347–366.

Silber, H., Stark, T., Blom, A., & Krosnick, J. (2019). Implementing a multi-national study of questionnaire design. In T. Johnson, B.-E. Pennel, I. Stoop, & B. Dorer (Eds.), *Advances in comparative survey methods: Multinational, multiregional, and multicultural contexts (*3MC*)* (pp. 161–180). Hoboken, NJ: Wiley.

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

Smyth, J. D., Dillman, D. A., Christian, L. M., & McBride, M. (2009). Open-ended questions in web surveys: Can increasing the size of answer boxes and providing extra verbal instructions improve response quality? *Public Opinion Quarterly*, *73*, 325–337.

Stark, T. H., Silber, H., Krosnick, J. A., Blom, A. G., Aoyagi, M., Belchior, A., ... Lawson, K. (2018). Generalization of classic question order effects across cultures. *Sociological Methods & Research*. doi: 10.1177/0049124117747304

Tourangeau, R., Conrad, F. G., Couper, M. P., & Ye, C. (2014). The effects of providing examples in survey questions. *Public Opinion Quarterly*, *78*, 100–125.

Triandis, H. C. (1988). Collectivism v. individualism: A reconceptualization of a basic concept in cross-cultural social psychology. In G. Verma & C. Bagley (Eds.), *Cross-cultural studies of personality, attitudes and cognition* (pp. 60–95). London, England: Palgrave Macmillan.

Triandis, H. C., Marín, G., Lisansky, J., & Betancourt, H. (1984). Simpatía as a cultural script of Hispanics. *Journal of Personality and Social Psychology*, *47*, 189–227.

Van de Vijver, F., & Chasiotis, A. (2010). Making methods meet: Mixed designs in cross-cultural research. In J. A. Harkness, M. Braun, B. Edwards, T. Johnson, L. Lyberg, P. Mohler, B.-E. Pennell, & & T. Smith (Eds.), *Survey methods in multinational, multiregional, and multicultural contexts* (pp. 455–473). Hoboken, NJ: Wiley.

Van Vaerenbergh, Y., & Thomas, T. D. (2012). Response styles in survey research: A literature review of antecedents, consequences, and remedies. *International Journal of Public Opinion Research*, *25*, 195–217.

Wells, T., Vidalon, M., & DiSogra, C. (2010). Differences in length of survey administration between Spanish-language and English-language survey respondents. In *JSM Proceedings, Survey Research Methods Section Proceedings of Survey Research Methods Section* (pp. 6186–6191). Alexandria, VA: American Statistical Association. Retrieved from http://www.asasrms.org/Proceedings/y2010f.html

Wierzbicka, A. (1991). *Cross-cultural pragmatics: The semantics of human interaction*. Berlin, Germany: Mouton de Gruyter.

Willis, G. B. (2004). *Cognitive interviewing: A tool for improving questionnaire design*. Thousand Oaks, CA: Sage.

Zaharna, R. S. (1995). Understanding cultural preferences of Arab communication patterns. *Public Relations Review*, *21*, 241–255.

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

Author Biographies

Katharina Meitinger is an assistant professor of methods and statistics at Utrecht University (the Netherlands). She holds a doctoral degree from the University of Mannheim (2016). Her current research interests include web probing, open-ended questions, visual design of questionnaires, and measurement invariance. Email: k.m.meitinger@uu.nl.

Dorothée Behr is a senior researcher at GESIS—Leibniz Institute for the Social Sciences, Mannheim (Germany). She holds a doctoral degree from the University of Mainz (2009). Her current research interests include web survey design and translation and comparability of cross-cultural questionnaires. Email: dorothee.behr@gesis.org.

Michael Braun is a senior project consultant at GESIS—Leibniz Institute for the Social Sciences and adjunct professor at the University of Mannheim (Germany). He has specialized in cross-cultural survey methodology and analysis. Email: michael.braun@gesis.org.