gesis Leibniz-Institut für Sozialwissenschaften

Web probing – implementing probing techniques from cognitive interviewing in web surveys with the goal to assess the validity of survey questions

Dorothée Behr, Katharina Meitinger, Michael Braun, Lars Kaczmirek

June 2017, Version 1.0

Abstract

Web probing – that is, the implementation of probing techniques from cognitive interviewing in web surveys with the goal to assess the validity of survey items – has recently found its way into the toolbox of (cross-cultural) survey methodologists. These guidelines present the origins of web probing, its developments, the current knowledge on its implementation, analysis possibilities and tips for the implementation of web probing in the cross-cultural context. These guidelines summarize the main findings from two research projects on web probing funded by the German Research Foundation (DFG). Wherever possible and existing, findings from other research groups supplement this overview.

Citation

Behr, D., Meitinger, K., Braun, M., & Kaczmirek, L. (2017). Web probing – implementing probing techniques from cognitive interviewing in web surveys with the goal to assess the validity of survey questions. Mannheim, GESIS – Leibniz-Institute for the Social Sciences (GESIS – Survey Guidelines).

DOI: 10.15465/gesis-sg_en_023

Acknowledgement: We wish to thank Cornelia Neuert and Cornelia Züll (both GESIS) for their helpful comments when revising this document.

1. Introduction

Good survey practice requires that questionnaires are pretested and evaluated before they are used in the field. Otherwise researchers run the risk that the data collected do not measure what they are supposed to measure. Among the most prominent pretesting methods are expert review, cognitive interviewing, and pilot-testing (e.g., Yan, Kreuter, & Tourangeau, 2012; see for an overview: Saris, 2012). In these guidelines, we will focus on a rather new – and still evolving – methodological extension of cognitive interviewing, namely web probing (sometimes also called online probing). According to Beatty and Willis (2007, p. 288), cognitive interviewing "entails administering draft survey questions while collecting additional verbal information *about* the survey responses, which is used to evaluate the quality of the response or to help determine whether the question is generating the information that its author intends." A prominent method in cognitive interviewing is probing whereby interviewers ask follow-up questions right after a closed-ended question of interest (embedded) or at the end of the survey (retrospective) to learn about respondents' cognitive processes, their ways of understanding certain terms, their ways of constructing their answers, etc. As with survey research in general, cognitive interviewing has started to embrace the possibilities of online research. This is where web probing has its origins. In a nutshell, web probing is the implementation of probing techniques from cognitive interviewing in web surveys with the goal to assess the validity of survey questions.

2. What is web probing?

In web probing, we ask, within the context of web surveys, open-ended questions as follow-ups to closed-ended questions. The follow-ups are called probes. These probes are modelled on cognitive interviewing probes and thus allow gaining insights into the answering and thought processes of respondents. Figure 1 provides an example of a probe (screen). After having answered the closed-ended question "How important is it that citizens may engage in acts of civil disobedience when they oppose government actions"? respondents receive the probe "What ideas do you associate with the phrase 'civil disobedience'? Please give examples."

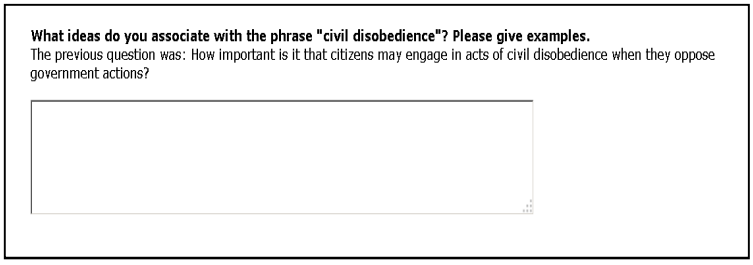

Figure 1: Example of a probe implemented in a web survey

Among the probe responses received for the above question were (in US-English; spelling not corrected):

- protests
- protest marches...non voilent / letter writing / picketing

- sit ins,nonviolent protests,some targeted remarks or barbs aimed at a politician without hatred or bias...i can remember a little of the watts riots from what my family showed me
- unsure
- I believe that non-violent disobedience is the only acceptable form, so that would include protesting unfair practices with marches and rallies to make sure the lawmakers were aware of the discontent.

The aim is to use the open-ended answers to examine whether the closed-ended questions measure what they are supposed to measure, that is, whether the closed-ended questions are valid. Moreover, if applied to a cross-national context, the qualitative data elicited through web probing allows checking for equivalence across countries. Different sources can be used to recruit respondents: probability-based panels, online access panels, crowdsourcing platforms or own resources (see section 3.8).

The GESIS research team consisting of Braun, Bandilla, Kaczmirek, Behr, & Meitinger, which has experimented with and tested probes in web surveys since 2010, have labeled their approach "web probing", thereby stressing mode and type of questions. The alternative term "online probing" is sometimes used to refer to the same concept. The following guidelines draw on and summarize much of the findings resulting from two research projects on web probing (CICOM and CICOM2, 2010-2015, Braun et al.)[1].

2.1 Origin: cognitive interviewing

As indicated in the introduction, web probing can be seen as a methodological extension of and supplement to cognitive interviewing. Even though cognitive interviewing has become best practice in survey research (Lenzner, Neuert, & Otto, 2016), there are limitations to the method, in particular the small sample sizes that are usually used and the associated danger to miss or overestimate errors or answer patterns (e.g., Conrad & Blair, 2009). In addition, the interactivity and flexibility that may be granted to cognitive interviewers can be regarded as a mixed blessing, depending on how far it reaches: On the one hand, interactivity and flexibility in probing allow following up on issues that have not been anticipated and that emerge during the interview itself. On the other hand, if several interviewers are used and granted too much flexibility and spontaneity in terms of probing, comparability of cognitive interviewing results may suffer through interviewer effects (Beatty & Willis, 2007; Conrad & Blair, 2009). Thus, it comes as no surprise that these limitations have provided the ground for extending cognitive interviewing to the web mode. Web surveys guarantee fast, easy, and wide-spread access to a large number of respondents. Furthermore, web surveys provide a medium for a standardized probing approach. Table 1 summarizes main advantages and disadvantages of cognitive interviewing vs. web probing (Edgar, Murphy, & Keating, 2016, as well as Meitinger and Behr, 2016, provide additional criteria such as cost or demography in their comparisons).

[1] Funding of the German Research Foundation (DFG) for "Optimizing Probing Procedures for Cross-National Web Surveys", (2012-2015, CICOM2, BR 908/5-1) and "Enhancing the Validity of Intercultural Comparative Surveys: The Use of Supplemental Probing Techniques in Internet Surveys" as part of SPP 1292: Survey Methodology (2010-2013, CICOM, BR 908/3-1).

Table 1: Comparative perspective on cognitive interviewing vs. web probing

		Cognitive interviewing	Web probing
Sample size of respondents	+		Large sample sizes & good assessment of prevalence of errors / patterns possible
	-	Typically small sample sizes	
Coverage of target groups	+	Special target groups, *including* illiterate, old, poor, ill, etc. can be reached	
	-		Only online population can be reached
Geographical coverage	+		Larger coverage as long as people have Internet access
	-	Typically limited to specific geographical areas	
Probing	+	Flexible, spontaneous probes possible, reacting towards unforeseen issues	Standardized probes → comparability
	-	If flexible and spontaneous approach prevails→ potential lack of comparability	Standardized probes → potentially insufficient information

Focusing on web probing, one can summarize the advantages and disadvantages of web probing as follows:

Among the advantages of web probing vis-à-vis cognitive interviewing are:

- ease of recruitment of large sample sizes;
- access to geographically/demographically diverse respondents, often with the possibility to quota-control the sample;
- elimination of interviewer effects thanks to standardized probing and an anonymous survey environment which reduces social desirability effects;
- no requirement of interviewers and thus no (additional) recruitment and training period required;
- the time needed for data collection is shorter;
- no need for transcriptions;
- in the cross-national context, relative ease of organizing a comparative study.

On the negative side, the disadvantages of web proving vis-à-vis cognitive interviewing include:

- its restrictions to population groups that can be reached online (e.g., via online access panels or crowdsourcing platforms) and that are sufficiently skilled in reading and writing;
- the lack of motivation by an interviewer and consequently an increase in probe nonresponse;
- the lack of interactivity, which would allow spontaneously acting on issues coming up in the probe response or which would allow rephrasing a probe that turns out to be problematic within the fielding context.

2.2 Related concept: crowdsourcing

Crowdsourcing "is the distribution of tasks to large groups of individuals via a flexible, open call" (Chandler & Shapiro, 2016, p. 54). Increased interest in crowdsourcing has led to the development of online labor markets (e.g., Amazon Mechanical Turk, MTurk, mainly including US citizens), and these allow academia to easily recruit convenience samples. Murphy, Keating, and Edgar (2013) and Edgar et

al. (2016), respectively, used crowdsourcing to recruit respondents for cognitive interviewing. They tested different implementation methods: self-administered cognitive interviewing with typed *and* spoken (audio) responses, and self-administered cognitive interviewing with typed responses only. The latter is essentially the same as web probing – asking probes online and receiving in return typed responses from the respondents. Murphy and colleagues stress the source of respondents, regardless of the pretesting implementation, which is why their different approaches are labeled "crowdsourcing."

2.3 Related concept: closed-ended probes

The research team around Scanlon (2016) has increased the scope of web probes to not only include open-ended probes but also targeted closed-ended probes. The closed-ended probes are used to take up issues identified during traditional cognitive interviewing. The goal is to be able to quantify cognitive interview findings in web surveys.

Figure 2 aims at systematizing the different developments that currently shape the evolving field of transferring techniques from cognitive interviewing to the online context. A distinction is made between *synchronous* and *asynchronous* communication. The former refers to a coordinated communication where an interviewer initiates and conducts a real-time interview; the latter refers to a communication that is based on an input (the web survey) that can be answered independently and at any time. Another distinction pertains to type of implementation, including different modes and mode combinations.

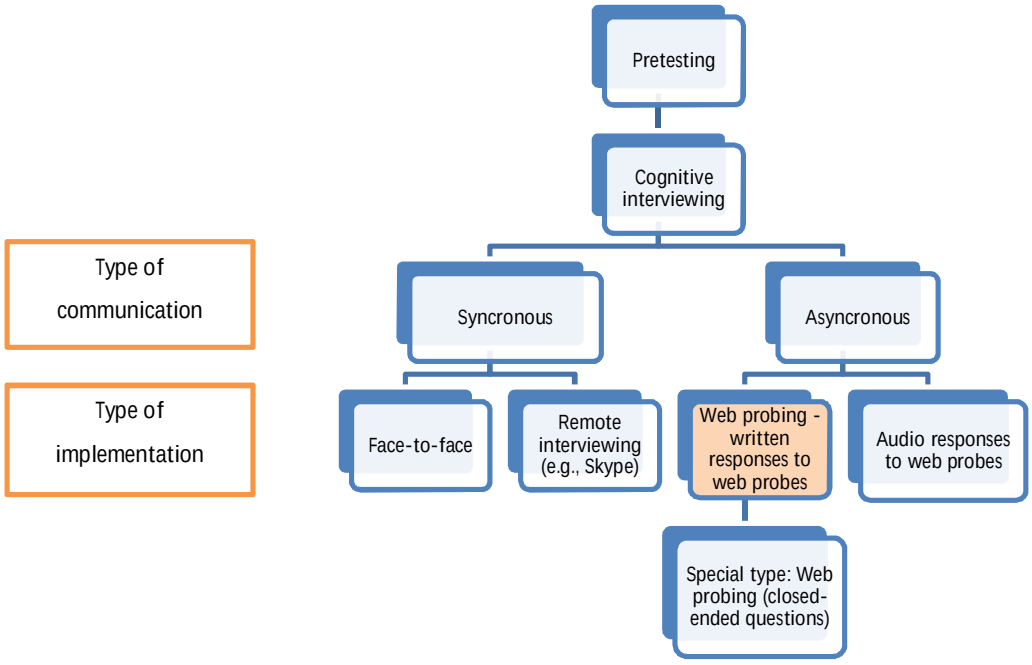

Figure 2: Web probing in the context of similar testing methods

2.4 Related concept: open-ended questions in general

Probes, as we understand them here, are a particular type of open-ended questions. They require narrative answers from the respondents, but *always* in relation to a foregoing closed-ended question. Thus, they differ from general open-ended questions that are asked *instead* of closed-ended questions e.g., "What are the most pressing problems in today's society?" (see Züll, 2016, or Couper, Kennedy, Conrad, & Tourangeau, 2011, on open-ended questions in general).

3. Implementation of probing in a web survey

As methodological research suggests, the implementation of open-ended questions in web surveys is not trivial. Open-ended questions are associated with a higher response burden than closed-ended questions, since pre-coded answers do not exist that could guide the respondents towards the things to consider in answering. Furthermore, typing the response itself may prove annoying or strenuous for some respondents. Additionally, a motivating interviewer is missing in the web mode so that everything needs to be designed in a manner that keeps the response burden low and makes the research intention as clear as possible. In the following, we summarize the main findings and experiences from the research projects conducted at GESIS, while stressing at the same time that much still needs to be learned. Existing research from other groups, where existing, supplements this overview.

3.1 Probe types

Web probing is an extension of cognitive interviewing and uses similar techniques. Different probe types can be used to address different issues. In the GESIS research projects, we have focused on:

- Category-selection probing: asking respondents for their reason(s) for having chosen a specific response category (e.g., "Please explain why you selected [chosen answer value; e.g. 'completely agree']." The responses usually take the form of a short argumentation. Category-selection probing is useful for checking whether the categories make sense to respondents, whether they are complete, distinct and allow enough differentiation or too much differentiation. Furthermore, category-selection probing can reveal silent misunderstandings of an item, that is, understandings that to the respondents make sense but which are not in line with the research goals.
- Comprehension probing: asking respondents to define how they understand a certain term, what ideas they associate with a certain term in general, or what a question is aiming at in a more abstract manner (e.g., "What do you consider to be a 'serious crime'?"). The responses typically take the form of a definition or a list of things (themes) that respondents think of in the context of the requested term. Comprehension probing is ideal for testing whether a term is understood as intended by the researcher.
- Specific probing: focusing on a particular detail of a term, on specific aspects that got activated in the context of a given question (e.g., "Which type of immigrants were you thinking of when you answered the question?") Once again, the responses, often short ones, typically contain a list of themes. Specific probes are very useful for getting an understanding of the breath that certain terms can have. For example, the term "immigrant" triggers associations of many different concepts such as specific countries or regions, and different attributes such as reasons to immigrate (asylum seekers), religious background, race, ethnicity, etc.

These probe types based on Willis (2005) and Prüfer and Rexroth (2005) worked well both in Germany and internationally in Canada, Denmark, Spain, Hungary, Mexico, the UK, and the US (countries in which we implemented web probing in our studies).

Other probe types and/or formulations can be viewed in Edgar et al. (2016) and in published presentations from the 2016 QDET2 conference[2]. Results from Edgar et al. (2016) suggest, however, that not all probes work equally well: The general probe "How did you arrive at the answer?" to follow up on clothing expenditure, for instance, did not provide as useful information when applied in the web mode when compared to traditional cognitive interviewing, at least in the described context. Added to this: In the context of cognitive interviewing, there is research that indicates that more specific probes are recommended over general probes (Foddy, 1998). This may be even truer in the web context. Probes need to clearly address the intended research goal, especially given that an intervening and possibly correcting interviewer is missing. It is also important to ask probes where respondents, in theory at least, can provide a somewhat longer answer; needless to say that one should avoid probes that only trigger yes/no responses; after all, in such situations the whole exercise of qualitative probing would be futile.

In sum, web probing hinges on the suitability of the chosen probe type and also on its specific formulation. The researcher has no chance to amend the probes during individual web survey sessions; therefore, the probe needs to be as targeted as possible and in line with the desired response format (see for similar requirements for open-ended questions in general, Züll, 2016).

In cross-cultural cognitive interviewing studies, current research is investigating which cognitive methods do not work across cultures in an equivalent fashion. Paraphrasing is one of those techniques where some groups of respondents fail to provide usable answers (Willis, 2015). Similar research crossing wider cultural boundaries is missing for cross-national web probing.

3.2 Probe placement

To disentangle the response process for the closed-ended questions from the probing process and thus to keep the 'usual' survey experience of closed-ended questions as stable as possible, we have implemented the probe on a separate screen following the respective closed-ended question (e.g., Behr, Kaczmirek, Bandilla, & Braun, 2012), as shown in Figure 3.

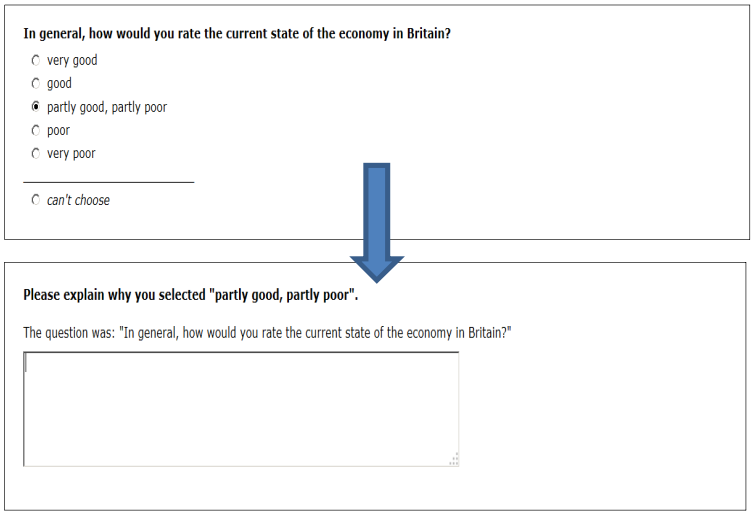

Figure 3: Closed-ended item and probe on next screen, making use of automatically inserting the previous answer into the next question

[2] Sessions "Web probing methods for pretesting" and "Web probing: considerations, uses, and practices"

However, if there are *many* probing questions in a survey, respondents might show signs of learning and start answering differently to closed-ended items due to their anticipation of an open-ended question and the required thought processes. First research by Fowler & Willis (2016) looked into whether respondents' answers to *closed*-ended questions differ depending on whether questions are probed immediately afterwards (embedded) vs. at the end of the survey (retrospective) – with inconclusive results. Couper (2013) finds some effects when a probe or 'commenting', as he calls it, is *systematically* implemented for a series of items in a survey (10 items of a scale in his case). This issue certainly should be followed up in research, that is, effects on closed-ended items depending on the number of probes and on whether probes appear on the same screen as the item, on a subsequent screen, or right at the end of a survey.

One may argue, at least as far as pretesting or post-hoc evaluation is concerned, that a slightly different response behavior to closed-ended items is irrelevant as long as the open-ended answers allow assessing the question's reliability and validity, and as long as relevant themes are put forward by respondents. Moreover, in the traditional cognitive interview the situation is similar: The respondents know right from the start that probes are going to follow, which may deepen or modify thought processes in comparison to the usual survey interview (Beatty & Willis, 2007; Conrad & Blair, 2009). For a main study implementation of web probing, care should be taken to reduce a potential impact on closed-ended items as much as possible (see section 4.2).

Beyond (possible) effects on *closed*-ended items, Fowler & Willis (2016) looked into the number of themes mentioned, depending on whether probes are embedded or asked retrospectively. Overall, similar themes are presented, albeit with a slight tendency to more relevant probe answers in the embedded condition.

3.3 Probe presentation

As mentioned before, we have asked the probe on a separate screen (see Figures 4 and 5). To alleviate response burden with such a design, respondents should be provided with the corresponding closed-ended item and, if relevant, their closed-ended answers on the probe screen. Thus, recall is aided, respondents can concentrate on the probe itself and probe nonresponse is decreased (Behr et al., 2012). For numerical scales, it is furthermore advisable to repeat the end labels of the scale to avoid confusion (see Figure 5).

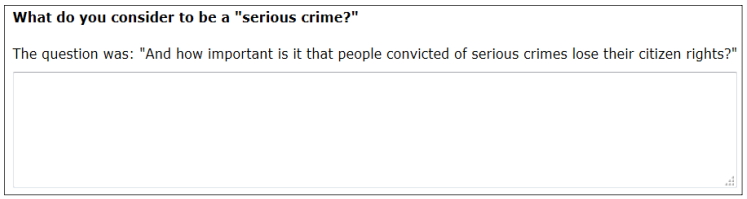

Figure 4: Specific probe

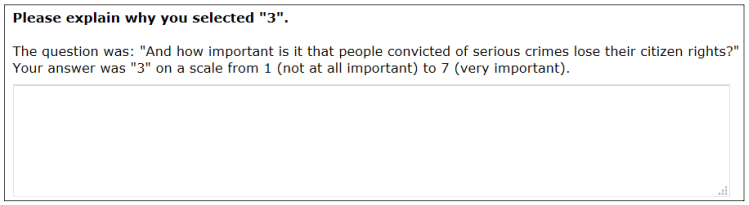

Figure 5: Category-selection probe

3.4 Sequence of probes

Regarding the sequence of probes, a researcher needs to make two decisions, first, regarding the sequence of probes referring to the same item, and second, regarding the sequence of probes across the entire survey.

Probes referring to the same item – the sequence of multiple probes

Researches may want to follow up on one item with several probes, because several aspects are worthwhile investigating, e.g. the reasons that lead to the selection of an answer option by the respondent and the understanding of a key term.

In an experimental study, we tested the sequences

- category-selection probe/specific probe/comprehension probe vs.
- comprehension probe/specific probe/category-selection probe.

We found that the design with category-selection probe coming first fared better. Overall, it increased response rates for the probes in total. It also decreased mismatching answers, that is, answers that do not fit to the probe type asked. Despite this general tendency, we have to issue a cautionary note: In our cross-national study where we tested the different sequences, the distribution of nonresponse and mismatching answers for this experiment was different across countries. These culture-specific effects call for further research into probe types and sequences in cross-cultural web probing (Meitinger, Braun, & Behr, 2017).

Probes across the entire survey – positioning of probes across the survey

In terms of the distribution of probes across the entire survey, we recommend to carefully reflect on the sequence and to design the probes in such a way that unintended habituation effects can be prevented: Behr et al. (2014) found that respondents habituate to a given probe type (e.g., category-selection probe), with its specific text-box size and overall layout, if this probe comes up repeatedly without interruption. What happened in their case was that after four or five repetitions, the *same visual outlook* of a subsequent but different probe type seems to have suggested to respondents that it is once again a category-selection probe. Rather than consciously reading the subsequent probe question, quite a large number of respondents answered in terms of their expectation and thus provided mismatching answers, which reduced the potential of data analysis. Therefore, with diverse probe types in a survey, the overall design (what is bold, size of text box, etc.) as well as the sequence of probe types should be considered so that mismatching answers, that is, those that do not fit the probe, can be kept to a minimum.

3.5 Number of probes

The sequence of probes is closely connected to the number of probes. We can report our project experiences, which go up to maximum 8 or 9 probes in a 15-minute survey – which seemed to work well. We usually took care to have the probes and a certain number of closed-ended items take turns, that is, there were no scales with many items which were probed one after the other. In case all items from a longer scale were relevant, we implemented splits to have only subgroups of respondents receiving a given set of probes, which then allowed testing much more items in a survey (Meitinger & Behr, 2016).

In our projects, respondents were informed on the existence of open-ended questions on the introduction screen (see below), but the survey was on purpose *not* framed as a pretest. Differently framed introductions (e.g., as a pretest) may have an effect on the number of probes one can ask. Also

the respondent source (online access panel, MTurk, etc., see section 3.8) may play a role concerning the willingness to answer probes. These issues require further investigation.

For illustration purposes, we gave the following information at the start of the survey.

> Throughout the survey, please take into account the following instructions:
>
> - There are no "right" or "wrong" answers to the questions. Please give the answer that best corresponds to your opinion.
> - The survey is unusual in that, for some questions, we ask you to provide reasons for your answer or to describe what you had in mind when answering.
> - Please take some time to answer these open questions. Your answers will help us to better understand the data which we collect in different countries.

3.6 Text box size

For open-ended questions, there is ample research that the text box-size gives cues to the respondents as to the desired answer depth, length or format. The same has been found for web probing (Behr, Bandilla, Kaczmirek, & Braun, 2014), and thus attention should be paid to fit the text box size to the desired answer type or format. For example, a desired explanation requires a larger text box than a question where researchers are only interested in, let's say, the types of sports activities that a respondent was considering when answering a question about sports. This means that a text box for a category-selection probe should be larger than a text box for a specific probe.

Guiding respondents in their answer behavior by adjusting the text box size to the probe type can also help to prevent mismatching answers, that is, answers that do not fit to the probe type.

3.7 Nonresponse reduction & tool support

Due to increased response burden, open-ended questions are particularly prone to nonresponse, and this is particularly the case in the web context. Cognitive probes are no exception in this regard. To reduce probe nonresponse, Kaczmirek, Meitinger, & Behr (2017) have developed a tool to automatically detect different types of probe nonresponse *during* the survey (see Table 2 for non-response types) and to follow up with a suitable follow-up probe and a tailored motivational statement. This tool, which is based on real-life corpora of existing probe nonresponse answers transformed into regular expressions, can be integrated into the survey environment; it is freely available in German, English, and Spanish, with corresponding information on its implementation and performance. For more information, please consult Kaczmirek et al. (2017).

Depending on the source of respondents (online panel, crowdsourcing platforms such as MTurk), nonresponse may or may not be an issue. Apparently, in MTurk nonresponse is not that big of a problem, since respondents are paid for a successfully performed service (Lee, 2015). But also in our online panel surveys in the GESIS research projects, we were, in general, positively surprised by the amount of useable answers. The mean nonresponse rate was 9% when calculated for 30 different questions in two surveys (Kaczmirek, Meitinger, & Behr, 2015). But it can also reach up to 30%, especially if lack of motivation is combined with high levels of lack of knowledge on or interest in a certain concept or topic (Behr et al., 2014).

The probe nonresponse conversion tool (Kaczmirek et al., 2017) can equally be used *after* data collection to automatically code nonresponse so that the remaining answers are quickly available for substantive analyses.

As a side note: (Some) online panel providers allow oversampling without incurring additional costs so that a certain degree of probe nonresponse can be compensated.

Table 2: Categories of nonresponse

Category	*Type of Probe Nonresponse*
Category 1	Complete nonresponse: respondent leaves a text box blank
Category 2	No useful answer: response is not a word e.g., "dfgjh"
Category 3	Don't know: e.g., "I have no idea," "DK," "I can't make up my mind"
Category 4	Refusal: e.g., "no comment," "see answer above"
Category 5	Other nonresponse: responses that are insufficient for substantive coding: e.g., "my personal experience," "it depends," "just do," "just what it is"[3]
Category 6	One word only: respondent just writes a single word, e.g., "economy"[3]
Category 7	Too fast response: respondent takes less than two seconds to answer

3.8 Access to respondents

Several ways to recruit respondents can be distinguished:

3.8.1 Probability-based panels

The gold standard of online research is certainly probability-based online panels because they provide a full coverage of the general population. In some European countries (the Netherlands, Germany, France, Norway, Sweden, others likely to follow), representative online panels have been set up that can be used by the research community (e.g., Blom et al., 2016); also in the US, probability-based panels are available (Callegaro et al., 2014). The cross-cultural web research endeavor is fostered by projects such as the Open-Probability-Based Panel Alliance (http://openpanelalliance.org/), which at the time of writing includes panels in the Netherlands, Germany, and the US. A fielding of items in probability-based panels requires a research proposal and a subsequent review process, which is why quick access and turnaround is often impossible. These panels may thus not be an option for (small-scale) and time-critical pretesting of survey items. Furthermore, these panels may not be open for pretesting at all but rather require pretested or functioning items in the first place.

3.8.2 Online access panels for national and international surveys

An online (access) panel is constituted of a group of respondents who have voluntarily signed up for taking part in surveys at regular intervals. Numerous online panels exist, and they vary widely in their quality and coverage. Information on panels is provided in the providers' answers to the *28 Questions to Help Buyers of Online Samples*[4], and in so-called panel books that typically provide a demographic overview of the respondents in the panel. Furthermore, several panels are certified by *ISO 26362:2009: Access panels in market, opinion and social research -- Vocabulary and service requirements.*

Examples of web probing using online access panels: Meitinger (2017), Behr et al. (2012), Behr et al. (2014) – essentially all research conducted in the two GESIS research projects referred to in these guidelines.

[3] The applicability of nonresponse categories 5 and 6 depends on the desired research interest.
[4] https://www.esomar.org/knowledge-and-standards/research-resources/28-questions-on-online-sampling.php (7/29/2016).

3.8.3 Crowdsourcing platforms

Research increasingly draws on crowdsourcing platforms to recruit participants for certain tasks (see section 2.2), that is, on online platforms where people who are willing to complete tasks are matched with people who request the completion of such tasks (Chandler & Shapiro, 2016). Examples include Amazon Mechanical Turk (MTurk) or TryMyUI, a remote usability testing service. These platforms are a preferred means for recruiting convenience samples, particularly in the US. Also Facebook can be used in a similar vein to recruit respondents, even though it does not share the typical 'labor environment' of these crowdsourcing platforms.

Examples of web probing using crowdsourcing platforms/Facebook: Edgar et al. (2016).

3.8.4 Own resources

Proprietary panels or an own respondent pool may equally be used for recruiting respondents. The US Census Bureau, for instances, manages a nonprobability "affinity" panel that can be used for research for the Census Bureau's own purposes (Childs, Clark Fobia, Holzberg, & Morales, 2016).

4. Stages of implementation and analysis potential

Both the introduction to these guidelines and Figure 2 advocate web probing as a pretesting technique to assess the quality of questions prior to fielding the main survey. We will argue that it can be used at the pretesting stage, but also as part of the main survey and, finally, as a post-hoc evaluation tool.

4.1 Pretesting stage

For its use during pretesting, different authors see different potential in web probing, especially when contrasted with traditional cognitive interviewing: Meitinger and Behr (2016) regard traditional cognitive interviewing as the method of choice for in-depth exploration of (new) questions, due to the possibility to follow up on probes and the interactive nature of the conversation. Web probing, in contrast, is advisable when researchers are interested in answer patterns and their prevalence, when the probe types to use are known (which presupposes a certain knowledge of the research topic and ideally relevant hypotheses), and when a certain geographical spread is needed. Their recommended sequence would thus have cognitive interviewing followed by web probing – if the budget allows for both. Examples where the GESIS Pretest Lab conducts web probing as the sole testing method (in line with the research questions and other contextual factors of the study) include pretests by Lenzner and Soiné (2014) as well as Meitinger, Neuert, Beitz, and Menold (2016).

Scanlon (2016) follow the same sequence, that is, cognitive interviewing followed by web probing, but their approach includes closed-ended probes in web surveys built on the basis of cognitive interviewing results. Their aim is to quantify the findings from cognitive interviewing.

Edgar et al. (2016) go the other way round, recommending to start with web probing ("crowdsourcing" in their terminology) for a rough overview of all kinds of issues, to continue with focused in-depth cognitive interviewing, and then ideally to have another round of web probing. Where resources are insufficient, web probing alone may be an option to solve at least some of the questionnaire design issues that may otherwise go unnoticed.

Turning to the international context, web probing at the pretesting stage may become interesting if cognitive interviewing is not viable (e.g., no cognitive labs in some countries, extensive training

required, etc.), or if cognitive interviewing can only be done in one or two countries, but should be supplemented in additional countries by means of web probing.

All these approaches show that there is not yet a generally recommended approach.

4.2 Main production stage

Where research draws on the web for the main data collection, one can easily imagine implementing probing at *selected* questions in the main survey itself. Given that probes increase response burden and might trigger a different response behavior to closed-ended items (see section 3.1), this should be done economically, however, at least when embedded in the survey rather than retrospectively. One could also control for possible effects by having only a split of respondents answering probes rather than all respondents. Response patterns across split conditions could then be compared. Already in the mid-1960s, Schuman acknowledged the benefits of probes – he conceived the model of "random probes" whereby splits of respondents receive a limited number of probes for selected questions (e.g., 10 probes per respondents). He argues that

> "[t]hrough qualitative and quantitative review of random probe responses the survey researcher has an opportunity to increase his own sensitivity to what his questions mean to actual respondents [...] In research in other cultures—and under some conditions in one's own culture—it forms a useful supplement to standard attitude survey methods". [Schuman, 1966, p. 222]

Thus, Schuman was one of the first proponents of the mixed-methods approach. This approach is now coming to the fore again; it is particularly useful in cross-cultural studies where the different contexts shape respondents' thoughts and interpretations of items (van de Vijver & Chasiotis, 2010).

4.3 Post-hoc evaluation

Web probing can equally be implemented post-hoc to shed light on existing survey data to (1) explain anomalies in the data or to (2) assess problematic questions in general. Post-hoc evaluation may be especially interesting for surveys with multiple waves or rounds, and where feedback is needed to take decisions on whether items should remain in a survey or not. The aforementioned GESIS research projects aimed in particular at assessing cross-national equivalence where analyses of data from the International Social Survey Program (ISSP), as a case in point, had revealed problematic statistical patterns for a country or several countries. One of our preconditions of using access panel data to shed light on ISSP data was to compare the distributions of closed-ended questions of the panel data to ISSP data. Only when problematic distributions could be replicated did we use the panel data to carefully explain what might have triggered equivalence problems in the ISSP data. When working with multiple item measures, measurement invariance tests using Multi-Group Confimatory Factor Analysis (Jöreskog, 1971) might be a further solution to compare one's own data with the replicated survey.

Behr et al. (2014) explored the meaning of civil disobedience in a cross-country perspective and thus explained distribution anomalies in the ISSP data. Meitinger (2017) combined web probing with quantitative measurement invariance tests to assess the equivalence of ISSP items measuring constructive patriotism and nationalism. By using web probing, Meitinger could locate a problematic item as identified with the measurement invariance tests and explain the reasons for the missing comparability: The translations of the item "social security benefits" had a varying lexical scope and many Mexican respondents silently misunderstood this term as referring to the security situation in the country (e.g., prevalence of crime).

In a more general manner, Braun, Behr, & Kaczmirek (2013) elucidated the meaning of "immigrants" in a cross-national perspective by probing attitude items on migration from the ISSP. While immigrants can more or less easily be translated into different languages, its interpretation in a cross-country perspective has repeatedly been questioned (see also Heath, Fisher, & Smith, 2005).

5. Use cases: Errors, themes, and response combinations

In the following, we present different uses of the probe answers. Probe answers can be used for detecting errors. In such a case, errors may be coded along the components of the response process, that is, comprehension, retrieval, judgment, and response (Tourangeau, Rips, & Rasinski, 2000). Errors can then include issues such as 'vague topic' and 'unclear question,' 'problematic term,' or 'information unavailable.' The error analysis can be based on already established error code schemes, such as the Question Appraisal System (Willis & Lessler, 1999) or the error coding schema by DeMaio and Landreth (2004).

Besides the error perspective, substantive themes may be the core interest of researchers. In such a case, a coding scheme will have to be developed (inductively, deductively or both ways) specifically for the item of interest. For instance, Meitinger and Behr (2016) looked into the meaning of "achievements in arts and literature" as appearing in "How proud are you of Germany in each of the following [...] its achievements in arts and literature?" (ISSP). The probe asked in the web survey was: "What particular achievements in the arts and literature did you have in mind when you were answering the question?" Answers were then coded along the broad themes of literature, music, performing arts, and visual arts. The Meitinger and Behr paper explicitly compares the error with the theme perspective as applied to both cognitive interviewing and web probing. (Similar comparisons between cognitive interviewing and various types of crowdsourced web probing studies can be found in Edgar et al., 2016).

Web probing can equally be implemented to shed light on response patterns for a combination of items that – in theory – should not be correlated or should not receive the same level of (dis-)agreement (e.g., items with reversed wording, opposing subdomains of an item, etc.). Thanks to the web mode, probes can automatically be triggered when a certain contradictory answer combination occurs. Behr, Braun, Kaczmirek, and Bandilla (2012) thus explain why two gender items, one traditionally slanted and one slanted towards an egalitarian position, both receive levels of disagreement or agreement, respectively, even though this is not in line with the original measurement goals (see for a cognitive interviewing study on seemingly contradictory response combinations, Campanelli, Gray, Blake, and Hope, 2016).

6. Analysis of probing data

The analysis of probing data is what requires most time. In the following, we describe, albeit briefly, a full-fledged analysis approach, including coding schema development, coding, and statistical analyses.[5] However, superficial insights for certain probe types might already be provided by automatic text

[5] For a detailed description of content analysis, please refer to Früh (2011, in German) or to Krippendorff (2013, in English) or Neuendorf (2017, in English) Also the GESIS Survey Guidelines on open-ended questions provide information on possible analysis approaches (Züll, 2016).

analysis tools, e.g. the visualization tool Wordle (http://www.wordle.net/) or the text analysis tool TAPoRware (http://taporware.ualberta.ca/).

Coding schema development can start from scratch based on the given probe answers and patterns that emerge from those (inductive development). A large enough sample of answers should be drawn upon for this purpose. Coding schema development may also start from theory and translate hypotheses on the research topic into codes (deductive development). A combination of both approaches is also possible.

Coding schemata include definitions of codes, coding rules (e.g. on what to exclude or include), and example answers for the different codes. The schemata should be set up in such a way that they can be consistently understood and followed by (several) coders not involved in coding schema development.

A finalization of a coding schema should occur only after a few trial runs (of coding small numbers of open answers). Trial runs help to uncover problems in the coding rules and potentially even in the codes.

Coders need to be trained on the final coding schema, by explaining it to them and by having several rounds of exercises and feedback. Once they master the coding schema, one coder can code the entire data set. Otherwise, several coders should consistently be trained on the coding.

A second coder should be employed on a random sample of probe answers to produce a basis for the establishment of intercoder reliability (e.g. 20 % out of 500 answers; the intercoder sample size depends, however, on the overall sample size and should increase with smaller sample sizes). We have found it helpful to code in Excel, which is a direct export format from the survey software we use, and to have 0-1 entries per column/code for indicating whether a code applies or not. The Excel file can easily be imported into a statistical software package so that these codes can be used in statistical analyses alongside the other data from the questionnaire. Alternatively, software that is tailor-made for qualitative data, including Atlas.ti (http://atlasti.com/) or MAXQDA (http://www.maxqda.com/), can be used.

7. Special case: Cross-cultural web probing

When applied to the cross-cultural or cross-national context, several additional issues need to be considered in probing and analysis. In such studies, open-ended responses are typically produced in diverse languages. This raises the question as to how to analyze these responses. Ideally, the research team consists of players from all cultures or countries involved so that coding and analysis can be done by native speakers in the original language versions. Alternatively, project members are sufficiently skilled in the languages and cultures needed and therefore can perform the coding and analysis on the basis of the original language versions. Finally, translation of coded answers can be performed so that project members can draw on these for coding and analysis; care should be taken though that translators do not change the message of the responses, e.g. by making unambiguous what is truly ambiguous in the original response, by rendering clear what is not intelligible, etc. Besides a close meaning transfer, which of course should adhere to the syntactical and grammatical target language requirements, commenting of cultural allusions, cultural facts, persons, etc. is deemed helpful to understand responses against their cultural backdrop. There should be an open communication channel to the translators so that remaining linguistic and cultural queries can be raised at any time.

If a coding scheme is developed inductively, that is, based on the respondents' answers, it is paramount that responses from all languages are taken into account so that the categories do not lean towards a specific language (and culture/country), at the expense of response patterns in other languages (and

cultures/countries). Further advice on translation, code schemas, and coding in a comparative context can be found in Behr (2015).

Cross-cultural web probing presupposes, of course, that all items, probes, motivational statements, error messages – all text that can appear online – is available in the different country languages. More on translation of questionnaires can be found in Behr, Braun, & Dorer (2017).

8. Conclusion

The 2016 International Conference on Questionnaire Design, Development, Evaluation, and Testing (QDET2) has shown, with its two sessions on web probing, that web probing has arrived in the questionnaire design community, and is increasingly used as a way to improve the questionnaire. Methodological studies testing its feasibility and comparing its outcome to more established methods are currently being conducted by research teams worldwide. There is still much to learn, e.g. with regard to the kind of items that can be assessed by web probing. Most of the research presented in the present guidelines is based on attitude items; but how does web probing perform when the items to be tested are factual or behavior items, for instance? There is also uncertainty as to what the minimum sample size is that allows saturation of themes or errors.

In sum, web probing is a valuable contribution to the methodological tool box of social science researchers. Nevertheless, besides its strengths, its limitations should be considered prior to any use: Use a method for what it can achieve. Don't overload a method with things for which it isn't made (see also d'Ardenne & Collins, 2016).

References

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

Behr, D. (2015). Translating answers to open-ended survey questions in cross-cultural research: A case study on the interplay between translation, coding, and analysis. *Field Methods*, 27, 284–299. doi: 0.1177/1525822X14553175.

Behr, D., Bandilla, W., Kaczmirek, L., & Braun, M. (2013). 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. doi: 0.1177/0894439313485203.

Behr, D., Braun, M., Kaczmirek, L., & Bandilla, W. (2013). Testing the validity of gender ideology items by implementing probing questions in web surveys. *Field Methods*, 25, 124–141. doi: /10.1177/1525822X12462525.

Behr, D., Braun, M., & Dorer, B. (2017). Measurement instruments in international surveys. *GESIS Survey Guidelines.* Mannheim, Germany: GESIS – Leibniz Institute for the Social Sciences. doi: 10.15465/gesis-sg_en_006

Behr, D., Braun, M., Kaczmirek, L., & Bandilla, W. (2012). Item comparability in cross-national surveys: Results from asking probing questions in cross-national web surveys about attitudes towards civil disobedience. *Quality & Quantity*, 1, 127–148. doi: 10.1007/s11135-012-9754-8.

Behr, D., Kaczmirek, L., Bandilla, W., & Braun, M. (2012). Asking probing questions in web surveys: Which factors have an impact on the quality of responses? *Social Science Computer Review*, 30, 487–498. doi: 0.1177/0894439311435305.

Blom, A. G., M. Bosnjak, A. Cornilleau, A.-S. Cousteaux, M. Das, S. Douhou, & U. Krieger (2016). A comparison of four probability-based online and mixed-mode panels in Europe. *Social Science Computer Review*, 34, 8–25. doi: 0.1177/0894439315574825.

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

Callegaro, M., Baker, R., Bethlehem, J., Göritz, A. S., Krosnick, J. A., & Lavrakas, P. J. (2014). Online panel research. History, concepts, applications, and a look at the future. In M. Callegaro, R. Baker, J. Bethlehem, A. S. Göritz, J. A. Krosnick, & P. J. Lavrakas, *Online panel research: A data quality perspective* (pp. 1–22). Wiley & Sons.

Campanelli, P., Gray, M., Blake, M., & Hope, S. (2016). Cognitive interviewing as tool for enhancing the accuracy of the interpretation of quantitative findings. *Quality & Quantity*, 50, 1021–1040. doi: 10.1007/s11135-015-0188-y.

Chandler, J, & Shapito, D. (2016). Conducting clinical research using crowdsourced convenience samples. *Annual Review of Clinical Psychology*, 12, 53–81. doi: 10.1146/annurev-clinpsy-021815-093623.

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

Couper, M. P. (2013). Research note: Reducing the threat of sensitive questions in online surveys? *Survey Methods: Insights from the Field*. Retrieved from http://surveyinsights.org/?p=1731.

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.

D'Ardenne, J., & Collins, D. (2016). Combining multiple evaluation methods: What does it mean when the data appear to conflict? International Conference on Questionnaire Design, Development, Evaluation, and Testing,

Miami, Florida. Retrieved from https://ww2.amstat.org/meetings/qdet2/OnlineProgram/Program.cfm?date=11-10-16

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

Edgar, J., Murphy, J., & Keating, M. (2016). Comparing traditional and crowdsourcing methods for pretesting survey questions. *SAGE Open*, October-December, 1–14. doi: 10.1177/2158244016671770.

Foddy, W. (1998). An empirical evaluation of in-depth probes used to pretest survey questions. *Sociological Methods Research*, 27, 103–133.

Fowler, S. L., & Willis, G. (2016). The practice of cognitive interviewing through web probing. International Conference on Questionnaire Design, Development, Evaluation, and Testing, Miami, Florida. Retrieved from https://ww2.amstat.org/meetings/qdet2/OnlineProgram/Program.cfm?date=11-11-16.

Früh, W. (2011). *Inhaltsanalyse - Theorie und Praxis* (Vol. 7., rev. ed.). Konstanz: UVK Medien.

Heath, A., Fisher, S., & Smith, S. (2005). The globalization of public opinion research. *Annual Review of Political Science*, 8, 297–333. doi: 10.1146/annurev.polisci.8.090203.103000.

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

Kaczmirek, L., Meitinger, K., & Behr, D. (2015). Item nonresponse in open-ended questions: Identification and reduction in web surveys. Conference of the European Survey Research Association, Reykjavik, Iceland. Retrieved from: http://www.europeansurveyresearch.org/conference/programme2015?sess=33#821.

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*. Cologne, 2017 (GESIS Papers 2017/01).

Krippendorff, K. (2013). *Content analysis. An introduction to its methodology*. Los Angeles: Sage Publication, Inc.

Lee, S. (2015). Personal communication on nonresponse in MTurk.

Lenzner, T., Neuert, C., & Otto, W. (2016). Cognitive Pretesting. *GESIS Survey Guidelines*. Mannheim, Germany: GESIS – Leibniz Institute for the Social Sciences. doi: 10.15465/gesis-sg_en_010.

Lenzner, T., & Soiné, H. (2014). German Internet Panel (GIP) – Module "Inflation" November Wave 2014. Cognitive Online-Pretest. GESIS Project Reports. Version: 1.0. GESIS - Pretestlab.

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

Meitinger, K., & Behr. D. (2016). Comparing cognitive interviewing and online probing: Do they find similar results? *Field Methods*, 28, 363–380. doi: 0.1177/1525822X15625866.

Meitinger, K., Braun, M., & Behr, D. (2017). Sequence matters in online probing: The impact of the order of probes on response quality, motivation of respondents, and answer content. *Submitted manuscript.*

Meitinger, K., Neuert, C., Beitz, C., & Menold, N. (2016). Pretesting of special module on ICT at work, working conditions & learning digital skills. Cognitive Online Pretest. GESIS Project Reports. Version: 1.0. GESIS - Pretestlab.

Childs, J., Clark Fobia, A., Holzberg, J. L., & Morales, G. (2016). A comparison of cognitive testing methods and sources: In-person versus online nonprobability and probability methods. International Conference on Questionnaire Design, Development, Evaluation, and Testing, Miami, Florida. Retrieved from https://ww2.amstat.org/meetings/qdet2/OnlineProgram/Program.cfm?date=11-11-16.

Murphy, J., Keating, M., & Edgar, J. (2013). Crowdsourcing in the cognitive interviewing process. Proceedings of the 2013 Federal Committee on Statistical Methodology (FCSM) Research Conference. Retrieved from https://fcsm.sites.usa.gov/files/2014/05/H1_Murphy_2013FCSM.pdf.

Neuendorf, K. A. (2017). *The content analysis guidebook*. Los Angeles: Sage Publication, Inc.

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

Saris, W. E. (2012). Discussion. Evaluation procedures for survey questions. *Journal of Official Statistics*, 28, 537–551.

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

Schuman, H. (1966). The random probe: A technique for evaluating the validity of closed questions. *American Sociological Review*, 31, 218–222.

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

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

Yan, T., Kreuter, F., & Tourangeau, R. (2012). Evaluating survey questions: A comparison of methods. *Journal of Official Statistics*, 28, 503–529.

Willis, G. B. (2005). *Cognitive interviewing: A tool for improving questionnaire design*. Thousand Oaks, CA: Sage.Willis, G. B. (2015). Research synthesis. The practice of cross-cultural cognitive interviewing. *Public Opinion Quarterly*, 79, 359–395.

Willis, G. B., & Lessler, J. T. (1999). *Question Appraisal System QAS-99*. Rockville, MD: Research Triangle Institute.

Züll, C. (2016). Open-Ended Questions. *GESIS Survey Guidelines*. Mannheim, Germany: GESIS – Leibniz Institute for the Social Sciences. doi: 10.15465/gesis-sg_en_002.