The International Journal for
Translation & Interpreting
Research
trans-int.org

Probing for sensitivity in translated survey questions: Differences in respondent feedback across cognitive probe types

Zeina Mneimneh[a]
zeinam@umich.edu

Kristen Cibelli Hibben[a]
kcibelli@umich.edu

Lisa Bilal[b,c,d]
lisabilal@gmail.com

Sanaa Hyder[b,c,d]
ssanaahyder@hotmail.com

Mona Shahab[b,e]
monakshahab@gmail.com

Abdulrahman Binmuammar[b,c,d]
muammar@kfshrc.edu.sa

Yasmin Altwaijri[b,c,d]
yasmint@kfshrc.edu.sa

[a] *Survey Methodology Program, Survey Research Center, University of Michigan*
[b] *King Salman Center for Disability Research, Riyadh, Saudi Arabia*
[c] *Biostatistics, Epidemiology and Scientific Computing Department, King Faisal Specialist Hospital and Research Centre, Riyadh, Saudi Arabia.*
[d] *SABIC Psychological Health Research & Applications Chair (SPHRAC), College of Medicine, King Saud University, Saudi Arabia*
[e] *Leiden University Medical Center, Leiden University, Leiden, The Netherlands*

DOI: 10.12807/ti.110202.2018.a06

Abstract: One of the core components of the TRAPD (Translation, Review, Adjudication, Pretesting and Documentation) team approach to translation in survey research is pretesting. Cognitive interviewing is increasingly being used for pretesting survey questionnaires adapted to different populations. Exploring the issue of question sensitivity is particularly relevant when adapting a questionnaire to a population different than the one for which it was designed. However, little guidance exists on the use of cognitive interviewing, and specifically, the types of verbal probes, to elicit respondent feedback on question sensitivity. In preparation for the Saudi National Mental Health Survey, cognitive interviewing was carried out to pretest the Arabic version of the World Mental Health survey instrument (CIDI 3.0). Different types of cognitive probes: proactive direct, proactive indirect and general probes were randomly assigned to survey questions to investigate differences in the feedback elicited by each type of probe. Findings suggest that different types of cognitive probes that are designed to explore perceived sensitivity of the survey questions elicit different amounts and types of feedback. An indirect cognitive probe identified a topic to be sensitive in more instances than a direct probe or a general probe. A general probe, on the other hand, elicited more non-codable feedback especially when paired with a survey question that asks about a more abstract concept such as the respondent's feelings.

Keywords: Translation, pre-testing, sensitive questions, cognitive interviewing

1. Introduction

When survey research requires content to be rendered in another language, a team approach known as TRAPD (translation, review, adjudication, pretesting and documentation) (Harkness, Villar, & Edwards, 2010) is recommended. While each step has its own significance, Willis and colleagues (2010) emphasize that no matter how well the earlier ones (translation, review, adjudication) are carried out, the "P" – Pretesting – step is essential, and is likely to find weaknesses in the translation. There are many aspects of a translated instrument that need to be tested. One important aspect is the level of perceived sensitivity of the survey questions. Both Lee (1993) and Barnett (1998) emphasize the contextual nature of question sensitivity and the importance of understanding how survey respondents think about sensitivity. Barnett (1998) goes on to recommend that the level of perceived question sensitivity be established prior to the study and during pilot work. One of the most powerful tools for pretesting translated survey instruments is cognitive interviewing (Goerman, 2006; Harkness et al., 2003; Potaka & Cochrane, 2004). Beatty (2004) defines cognitive interviewing as:

> [the] practice of administering a survey questionnaire while collecting additional information about the survey response; this additional information is used to evaluate the quality of the response or to help determine whether the question is generating the sort of information that its author intends (p. 45).

This paper examines the perceived survey question sensitivity through the application of cognitive pretesting to a set of translated questions on mental health. We focus specifically on comparing the effectiveness of different cognitive follow-up verbal probes in uncovering the sensitivity of the translated survey questions and identifying the probe or probes that are most effective at encouraging respondents to discuss the perceived sensitivity of the translated survey items. In doing so we hope to guide practitioners in selecting appropriate probes for testing translated instruments and emphasize the general need for further empirical studies that evaluate the effectiveness of alternate probing models and types as concluded by Willis (2015a).

There is growing literature that discusses the application of cognitive interviewing in cross-cultural research for uncovering problems with survey items when they are translated for a different target language and culture (Agans, Deeb-Sossa, & Kalsbeek, 2006; Miller et al., 2011; Willis, 2015a; Willis, Kudela et al., 2008; Willis & Miller, 2011). When certain dimensions of survey items vary across cultures, such pretesting is essential in determining whether a translated instrument is well adapted and culturally relevant to its intended respondents.

An important culture-dependent dimension of survey questions is their level of perceived sensitivity (Andreenkova, forthcoming). Tourangeau and Yan (2007) define sensitive questions as a broad category of questions including those which trigger social desirability concerns, are seen as intrusive by respondents, or raise concerns about the repercussions of disclosing the information.

However, cultures differ in their views of the self and what is considered to be favourable, personal, or intrusive, and this may cause systematic variation in self-presentation. For example, western individualist cultures make salient norms of self-confidence and self-enhancement while Confucian-based collectivist cultures in East Asia focus on harmony, modesty, and fitting-in (Heine, 2007; Kitayama, Duffy, & Uchida 2007; Markus & Kitayama, 1991; Suzuki & Yamagishi, 2004; Yamaguchi, 1994). On the other hand, African, Latin American, Mediterranean, and Middle Eastern honour-based collectivist societies focus on the need to protect and maintain honour through the protection of social image and a positive presentation of the self and in-group members and a negative representation of the out-group (Uskul, Oyserman, & Schwarz, 2010).

Such cultural differences in social presentation can affect how respondents cognitively process the survey question and potentially edit the retrieved information to form a culturally "acceptable" response, especially in face to face interviews. It is therefore important to identify the culture's perceived sensitivity to a translated survey question so that design modifications may be made to the question itself, the instructions around the question, or the mode of the survey, so as to encourage honest responses and reduce potential biases during the implementation phase. A key question is how can researchers uncover the perceived sensitive nature of the requested information in a specific culture? Are there type(s) of cognitive probes that are better at uncovering the perceived sensitivity of a translated survey question?

One common technique used in cognitive interviewing to pretest a translated survey question and collect additional verbal information about the survey question itself is 'verbal probing'[1]. In this approach, which seeks to elicit specific information relevant to the survey questions (such as question comprehension), interviewers can ask scripted 'anticipated' probes or unscripted 'spontaneous' probes to address pre-identified concerns (Willis, 2005, p.88). When particular problems with a translated question are anticipated, verbal probing and, specifically, proactive probing can – when used judiciously[2] – be beneficial. As Willis (2005) argues, proactive probing can be considered more systematic because it is based on hypotheses about suspected flaws in the questions being evaluated. Further, using anticipated or scripted probes can help minimize the variance between interviewers since they use the same probes and may be particularly advantageous when interviewers are relatively inexperienced.

There are several examples of scripted probes in the literature (see Willis, 2005), and some are related to understanding the perceived sensitive nature of the survey question. Examples in the literature include: 1) *direct probes* about how the respondent him/herself reacts to the question, such as "How did you react to being asked this question?" (Vernon, 2005), and "How easy or difficult was it for you to come up with an answer?" (Berrigan et al., 2010), and 2) *indirect probes,* asking the respondent how he/she thinks other people would react to the survey question, such as "In general, do you feel that people might purposely say … or would they try to answer accurately?" (Warnecke et al, 1997), and "How do you think people will react to being asked about …?" (Vernon, 2005). While these types of probes have been used previously, their relative effectiveness in eliciting a respondent's feedback about question sensitivity has not been empirically tested[3]. Thus, researchers who are interested in cognitively testing whether the translated survey questions are perceived as sensitive in the desired culture have no specific guidance on which scripted probes are better suited for that purpose.

Since the feedback given by the respondent is heavily dependent on the probes the interviewer uses, understanding which probes are more effective is essential. This is especially important if translated questions designed for one

[1] For a more detailed discussion on verbal probing, see Willis (2005).

[2] As Conrad and Blair (1996, 2001) have noted, a possible downside to probing, and the use of proactive probes in particular (especially if overdone by interviewers), is that searching for particular problems may create the appearance of problems that do not really exist.

[3] While Vernon (2005) used different types of verbal probes to examine respondents' feedback to potentially sensitive questions (about income and weight), the author did not randomize the probes and the objective of the paper was not to empirically test differences in respondent feedback between the two verbal probes.

culture are to be administered in another, in which perceived sensitivity might vary and where social desirability (either in general or its sub-types, such as impression management) might be relatively higher, such as in collectivist cultures (as compared to individualistic societies) (Bernardi, 2006; Bond & Smith, 1996; Lalwani, Shavitt, & Johnson, 2006; Triandis, 1995).

Consequently, this study compares the outcomes of three types of scripted probes used to elicit information about respondents' perceived sensitivity to a set of translated questions – proactive direct, proactive indirect, and general probe. Proactive *direct* probes (e.g., "How difficult is it to talk about this subject?") ask the respondents directly whether *they* find the subject difficult to talk about. Proactive *indirect* probes (e.g., "To what extent others would find it uncomfortable to talk about this issue?) ask the respondents indirectly about question sensitivity by inquiring whether or not *others* might find it uncomfortable to talk about the topic. Finally, general probes (e.g., "Tell me more about your opinion of this question?) ask respondents to elaborate in general about the question. The rationale for comparing these different types of probes is whether, similar to survey interviewing, asking respondents about a sensitive question related to themselves could lead to different feedback than asking them about a sensitive topic related to 'others'. Since respondents can distance themselves from the sensitive behaviour or attitude, researchers believe that respondents might be more likely to report that 'others' (including friends) engage in such sensitive behaviour or have such attitudes compared to reporting about themselves. This technique has been referred to as the nominative technique (Blair et al., 1977; Lee, 1993). This study investigates whether this phenomenon may also be observed in cognitive interviews that are used for testing translated instruments, and whether it uncovers higher rates of perceived sensitivity in a collectivist culture where talking about sensitive topics is generally avoided. Moreover, the study compares respondents' feedback from two proactive probes (direct and indirect) to a general probe that does not allude to question sensitivity and where respondents could intentionally or unintentionally avoid addressing the sensitive nature of a question.

To explore these issues, the probes - proactive direct, proactive indirect, and a general probe - were randomly assigned to a series of mental health questions in the Kingdom of Saudi Arabia (KSA), an honour-based collectivist culture. Identifying whether different cognitive probes could uncover different respondent views on the sensitivity of questions was important for pretesting the instrument that was translated from an original English questionnaire designed for a different culture. Findings from this experiment could guide practitioners who are translating survey questions and are concerned about the perceived sensitivity of the questions in the target population.

2. Methods

2.1 Background

Cognitive interviews used in this study were designed to pretest the adapted Arabic questionnaire of the Saudi National Mental Health Survey (SNMHS) in KSA. The SNMHS is part of the World Mental Health (WMH) Initiative that currently includes more than 30 national surveys across the globe (http://www.hcp.med.harvard.edu/wmh). The questionnaire used in all WMH surveys is the Composite International Diagnostic Interview (CIDI) 3.0. The CIDI 3.0 is a comprehensive and fully-structured interview designed to be used by trained lay interviewers for the assessment of mental disorders (Kessler & Ustun, 2004). Its original source language is English but the instrument has been translated into many languages.

The TRAPD translation model was implemented to carry out the Saudi adaptation of the CIDI. During the adaptation process several survey questions were selected for cognitive testing to explore their potential sensitivity in the Saudi culture. A sample of the selected questions and their translation is listed in the Appendix. The SNMHS team judged the question topics to be potentially sensitive in the Saudi culture based on their understanding of the prevailing culture and their expertise and experiences with perceptions of mental and physical health symptoms and social issues within the Saudi population. However, and as discussed by Barnett (1998), it is important to understand the perceived sensitivity of the question from the respondent's point of view. Barnett recommends that some kind of pilot work is needed to establish question sensitivity in relation to the population being sampled.

2.2 Cognitive probes

In the absence of empirical guidance on the type of probe that could encourage respondents to talk about the sensitivity of questions in a culture different than the one the questions were designed for, the authors decided to randomize three different probes across respondents: proactive direct, proactive indirect, and general probes. These probes were asked concurrently right after the respondent provides his/her answer to the survey item selected for testing. Table 1 presents the probe wording and the number of instances each probe was asked across 49 respondents.

Table 1: Probe wording and instances*

Probe	Probe Wording	Instances
Proactive Direct	English Wording: How difficult is it to talk about this issue? Arabic Wording: ما مدى صعوبة التحدث عن هذا الموضوع ؟	56
Proactive indirect	English Wording: To what extent would others find it uncomfortable to talk about this issue in such a survey? Arabic Wording: الى أي درجة تعتقد أن الناس قد تشعر بعدم الارتياح للتحدث عن هذا الموضوع في مقابلة مثل هذه؟	47
General	English Wording: Tell me more about your opinion of [how you feel about] this question? Arabic Wording: خبرني اكثر عن رأيك (ما شعورك اتجاه هذا) بهذا السؤال؟	34
	Total	**137**

*Instance: The number of times a probe was asked across all cognitive interviews

A purposive sample of Saudis was selected to participate in the cognitive interviews. It included both sexes of different age groups and educational backgrounds. Some of the respondents had missing data on their demographics (especially educational level). The total number of respondents with complete demographic information was 43. Mean age of these respondents was 33 with a range of 16-51 years old. Slightly more than half (53%) were males, 28% had less than high school education, 39% had some college education, 12% had an undergraduate college degree, and 21% had a graduate degree. Moreover, the sample included both, participants with a history of mental disorders (49%) and those without such history (51%). The former group was selected from a local Saudi clinic in Riyadh based on their diagnosis of certain mental health disorders such as obsessive compulsive disorder, mania, generalized anxiety disorder, social phobia, panic disorder, depression, and bipolar disorder. All respondents were informed that the objective of the study was to pretest a survey instrument that was adapted to Arabic. Informed consent was obtained from all participants.

The cognitive interviews were conducted by thirteen local Saudi interviewers, who were all broadly familiar with research methodology, owing to their varied backgrounds – in psychiatry, epidemiology, social work or other health sciences. These interviewers underwent a 5-day web-training session held by collaborators at University of Michigan with a local Saudi trainer facilitating the sessions in Riyadh. The training covered several topics including introduction to cognitive interviewing, types of cognitive interviewing techniques (including probing), how cognitive interviews are conducted, analysed and documented. The training also included a practical component where interviewers practiced carrying out cognitive interviews and asking probes in pairs.

The cognitive interviews were conducted either at the clinic where the patients (with a history of mental disorder) were recruited from, at the patients' houses, or at the research centre which the authors are affiliated with. While the interviewer conducted the interview and took some notes, an observer made additional notes as recording is culturally unacceptable.

Two bilingual coders independently read all the notes and coded respondents' feedback to each probe into one of three categories: the respondent thought the question was sensitive, not sensitive, or could not tell (i.e. the feedback was not codable) from the notes provided by the interviewer and the observer. The initial agreement rate between the two coders is 66%. Discrepancies between the coders were reviewed and resolved by two of the authors (one bi-lingual).

2.3 Research questions and analysis

Three main research questions were investigated:

Research Question 1: How do proactive direct, proactive indirect and general probes differ in the amount of feedback they elicit?

Research Question 2: How do proactive direct, proactive indirect and general probes differ in the content of feedback they elicit regarding the sensitive nature of the survey question?

Research Question 3: Does the content of the respondent feedback elicited by the different probes differ by the topic of the survey question being tested (i.e. survey questions asking about behaviours vs. feelings)[4]?

To investigate these research questions, first the association between respondent characteristics and respondent feedback across all the cognitive probes was tested. The aim was to confirm that the randomization of the different types of probes across respondents was successful and thus differences in respondent feedback is not driven by respondent characteristics. To test this, three ratios were calculated for each respondent: ratio of probes that generated non-codable feedback, ratio of probes that generated sensitive feedback, and ratio of probes that generated non-sensitive feedback. Each of these ratios was regressed (using multivariate linear regression models) on respondent's sex, age, education level and history of mental health disorder. None of the associations was statistically significant providing support that respondent-level characteristics were not related

[4] The different probes are the proactive direct, proactive indirect, and general (listed in Table 1). These probes were administered right after several survey questions to elicit respondent feedback on the perceived sensitivity of the survey question. A sample of the survey questions (asking about behaviour, attitudes, or feelings) are presented in the Appendix.

to the feedback outcomes. The rest of the analyses focus on the type of probes and the feedback outcomes (amount and content).

To investigate the first research question, the length of feedback was grouped into four categories: one word, a phrase (between 2 and 5 words), a sentence (between 6 and 10 words), and longer than a sentence (more than 11 words). For each of the probe types (direct proactive, indirect proactive, and general) the percentage of feedback that fell into each of these categories was calculated. For the second research question, the percentage of non-codable feedback, non-sensitive feedback, and sensitive feedback was also calculated for each type of probes. Differences in percentages across the probes were tested using Chi-square test statistics. All analyses were conducted using SAS 9.4 (SAS Institute, NC).

3. Results

3.1 Research question 1: Variation in respondent feedback length by type of probe

Table 2 summarizes whether the length of the feedback elicited by the respondent (one word only, a phrase, a sentence, or more than a sentence) varied by type of cognitive probe. Asking respondents a general probe produced lengthier responses than the proactive direct and proactive indirect probe; 29% vs. 19% vs. 16% of the probes respectively, elicited respondent feedback that is longer than a sentence. The direct probe elicited the briefest amount of feedback; where 20% of the direct probes elicited only one word, vs. 4% of the proactive indirect probes, and 9% of the general probe.

Table 2: Percentage of feedback consisting of one word, phrase, sentence or >sentence by probe type

Probe Type	Word Count			
	One word	Phrase	Sentence	> Sentence
Proactive Direct: How Difficult	19.6%	42.9%	21.4%	16.1%
Proactive Indirect: How others feel	4.3%	27.7%	48.9%	19.1%
General: How R Feels/Opinion	8.8%	29.4%	32.4%	29.4%
Chi-Square test (p-value)	15.21 (0.0187)			

3.2 Research question 2: Variation in respondent feedback about the sensitive nature of the survey question by probe type

Not only did the amount of feedback differ by probe type but the content of the feedback differed too. Findings presented in Table 3 suggest that asking respondents a proactive *indirect* probe (whether others find it uncomfortable to talk about the topic of the question) led to more feedback that identified the topic as sensitive, 74% of the feedback, compared to a direct probe, 34%, or a general probe, 21%. Asking respondents a direct probe (whether the respondent finds the topic difficult to talk about) on the other hand identified the topic as more non-sensitive, 52% of the instances, compared to either of the two other probes, 23% (for the indirect probe) and 35% (for the general probe) of the instances. Whereas, a general probe (how the respondent feels about the topic of the question) produced the highest non-codable feedback, 44% compared to the other two probes 14% (direct probe) and 2% (indirect probe).

Table 3: Percentage of probes eliciting “Not codable”, “Not sensitive”, and “Sensitive” feedback by probe type

Probe Type		Feedback Outcome	
	Not codable	Not sensitive	Sensitive
Proactive Direct: How Difficult (N=56)	14.3%	51.8%	33.9%
Proactive Indirect: Others Find it Uncomfortable (N=47)	2.1%	23.4%	74.5%
General: How R Feels/Opinion (N=34)	44.1%	35.3%	20.6%
Chi-Square test (p-value)	41.18 (<0.001)		

3.3 Research question 3: Variation in respondent feedback by probe type by survey question type

The overall differences in the content of feedback observed between probes seem to hold for the survey questions asking about feelings (such as worthlessness). The proactive indirect probe elicited the highest percentage of feedback identifying a survey question about feelings as sensitive, 76% compared to 23% and 13% for proactive direct probe and general probe respectively. The proactive direct probe on the other hand elicited the highest percentage of feedback identifying a feeling question as non-sensitive and the general probe generated the highest percentage of feedback that is not codable (Table 4, Questions about Feelings). However, for questions asking about behaviours (such as using a weapon), while the proactive indirect probe led to more feedback that identified the behaviour as sensitive (73%), the general probe elicited more feedback that finds the topic to be not sensitive (45%), and there was a small difference in the percentage of feedback that was not codable between the proactive direct and the general probe (14% vs. 18%). However, the sample was not large enough to detect significant differences when restricting the analysis to the survey questions asking about behaviours (Table 4, Questions about Behaviours).

Table 4: Percentage of probes eliciting “Not codable”, “Not sensitive”, and “Sensitive” feedback by probe type for questions about behaviours and questions about feelings

Probe Type		Feedback Outcome	
	Not codable	Not sensitive	Sensitive
Questions about Behaviours			
Proactive Direct: How Difficult (N= 22)	13.6%	36.4%	50.0%
Proactive Indirect: Others Find it Uncomfortable (N= 22)	4.6%	22.7%	72.7%
General: How R Feels/Opinion (N= 11)	18.2%	45.4%	36.4%
Chi-Square test (p-value)	----[a]		
Questions about Feelings			
Proactive Direct: How Difficult (N= 34)	14.7%	61.8%	23.5%
Proactive Indirect: Others Find it Uncomfortable (N= 25)	0.0%	24.0%	76.0%
General: How R Feels/Opinion (N= 23)	56.6%	30.4%	13.0%
Chi-Square test (p-value)	40.40 (0.001)		

[a] 44% of cells have expected counts less than 5 and thus Fisher exact test is used. P-value=0.2730

4. Discussion

Testing translated survey questions in the target population before field administration is a critical component of the TRAPD translation approach. One dimension of a survey question that is context and culture dependent is its perceived level of sensitivity. Identifying the most effective pre-testing probe that encourages respondents to discuss the perceived sensitive nature of the translated survey item is essential. This study found that different types of cognitive probes that are designed to explore the sensitive nature of translated survey questions elicit different amounts and types of sensitive feedback. A cognitive probe asking whether 'others' would find the topic of the question uncomfortable to talk about – a proactive indirect probe – identified a topic to be sensitive in more instances than a direct probe that asks about the respondent him/herself, or a general probe that does not allude to the sensitivity of the topic at all (e.g. how the respondent feels about the question in general). This difference was consistent across both types of the tested survey questions, i.e. those asking about behaviours and those about feelings. This is in line with the nominative technique, which theorises that indirect questions that ask about the sensitive behaviour of others, including friends, would elicit more sensitive responses than direct questions that ask about the respondent him/herself (Barnett, 1998; Lee, 1993; Tourangeau & Yan, 2007). When asked about others, respondents can distance themselves from the behavior, judge it more objectively and potentially report a more honest opinion. In cognitive interviews, expressing that a question is sensitive to the interviewer could allude to the sensitive nature of the answer - i.e., 'I engage in this behavior and that's why I find it sensitive'. However, expressing that a question is not sensitive could indicate that 'I don't engage in this behavior and thus there is no reason why I find it sensitive'. Thus, in cognitive interviewing, when asked directly about the perceived sensitive nature of a survey question, a respondent might deny it so as to avoid any social stigma. This phenomenon could be more common in collectivist societies, where social conformity has been shown to be higher than in individualistic societies (Bernardi, 2006; Bond & Smith, 1996; Lalwani et al., 2006; Triandis, 1995). These findings are also consistent with Storti's (1999) theory which posits that Middle Eastern countries fall somewhere in between the continuum of directness and indirectness with regard to expressing opinions; the Saudi respondents seemed to be open about sensitivity when a question was directed at others, while they were less open when the probe was directed to them personally. This suggests that indirect probes may be more effective than direct probes or general probes for eliciting information about question sensitivity in cultures that lie closer to indirectness on the continuum of communication styles. This could inform the practice of administering cognitive probes in similar cultures and potentially the phrasing of survey questions as translated for use in such cultures. It is also possible, however, that the indirect proactive probe might have led respondents to exaggerate the sensitive nature of the question indicating a potential issue with the question that does not really exist. Although we do not believe that this is the case – as the survey questions selected for testing were judged by the local survey team to be potentially sensitive in this specific culture –, this potential over-reporting still needs to be empirically tested.

On the other hand, asking respondents how 'others' feel towards a topic might be measuring respondents' perception of the social norm surrounding the topic based on respondents' own beliefs, their behaviours, and interpretation of the term 'others'. Such a perception could be tapping into a different aspect of sensitivity than the one measuring respondents' difficulty in talking about the topic (through the direct probe). The distinction between the two aspects of sensitivity might be larger in a collectivist culture where 'others' might refer to out-group members who are seen to be critical of the respondent's behaviour.

It is important to note that the phrasing of the direct and the indirect probes in this study varied on two dimensions: directness (self vs. others), and the feeling associated with the topic (difficulty talking about the topic vs. the discomfort talking about the topic). Thus, differences in the feedback between the two probes are attributed to both dimensions rather than just the directness of the probe.

While all three types of probes elicited some non-codable feedback, the highest rate was for the general probe – "How do you feel about the question?", and more specifically for the survey questions that ask about a more abstract concept such as feelings. For the majority of the non-codable feedback, the respondent either misunderstood the probe that was administered after the survey question and elaborated on their answer to the survey question itself, or answered the probe but did not give enough information to judge if they think the content of the question is sensitive or not. For example, the respondent might reply "this situation that is mentioned in the question could happen to people with mental disorder." This is because the probe was not specific enough (as intended) and the interviewer, even though instructed to do so, did not follow up with additional *general* probes such as "we are referring to the topic of the survey question, how you feel about the topic itself". The high rate of non-codable feedback to the general probe could indicate respondents' difficulty with this type of probe as reported by other authors (see Willis, 2015b for a review) and more specifically when this general probe is paired with a survey question asking about an abstract concept. Still, such a high rate also raises the question of whether the interviewers had the needed skills to follow-up with unscripted non-leading general probes when the respondent did not provide the desired feedback. It is possible that the duration of the training, specifically the practice component, was not enough for interviewers who were new to the practice of cognitive interviewing. This highlights the need for longer training for inexperienced interviewers, close to real-time review of cognitive interviewing notes, and the implementation of more interviews in case some turned out to be ineffective, as discussed by Miller (forthcoming).

The present study provides some guidance to prospective projects relying on cognitive interviews to pretest the sensitivity of translated survey questions. However, this study has a few limitations. First, none of the interviewers had previous experience in conducting cognitive interviews and might have benefited from additional training. Second, the high rate of non-codable feedback for the general probe limits the accurate assessment of the feedback content (regarding topic sensitivity). However, if we apply the ratio of sensitive vs. non-sensitive codes among codable feedback to the non-codable feedback, the topic sensitivity rate for the general probe will still be lower than the indirect proactive probe and would be closer to the direct proactive probe. Third, the phrasing of the direct probe did not explicitly specify the agent (i.e. "How difficult it is for you to talk about this subject?". The reference was rather implicit "How difficult it is to talk about this subject". However, cognitive interviewers were trained to convey the meaning of all probes in a non-leading manner if they felt that the respondent misinterpreted the intended meaning. Still, observed differences in respondents' feedback between the scripted pro-active probes that are due to the directness of the probe could be an underestimate and an explicit reference to the agent in the phrasing might show higher differences. On the other hand, the phrasing of the indirect probe differed on both, directness of the probe and the feeling associated with talking about the topic. This makes it difficult to isolate the effect of directness only. Future research that disentangles this confounding and explicitly specifies the agent in each of the probes is needed if the objective is to evaluate the directness component only. Fourth, while having an observer (note taker) is a common practice when cognitive interviews are not recorded, it is possible that the observer's presence may have affected respondents' answers to the probes. However, this was consistent across the different types of probes and thus should not contribute to the observed feedback differences. Fifth, the sample size for cognitive interviewing is typically small, which might have led to lower statistical power to detect some of the differences especially those related to specific types of questions (such as questions asking about feelings). Finally, this experiment was conducted in a specific culture where the belief system is deeply grounded in religion and traditions. Future replication of such experiments on different cultural groups would enhance our understanding of the effectiveness of different types of cognitive probes and their generalization to different populations.

In conclusion, this is the first study that attempts to identify the most effective cognitive verbal probe to elicit feedback differences related to the perceived level of sensitivity of translated survey questions. The results of this study could inform practitioners as well as researchers interested in testing or measuring the sensitive nature of survey questions translated to a different culture. Utilizing an effective probe that uncovers the perceived sensitivity of a translated question has direct implications on the adaptation of the question. If a question is found to be sensitive in the target culture, design changes could be implemented and further testing could be done. Such design changes could be related to the question phrasing, respondent instructions or the mode of administration that would encourage accurate reporting and improve the quality of the data collected. Finally, and more generally, experiments designed to systematically assess differences in feedback elicited by different probes are needed. Such systematic assessment would inform questionnaire design decisions related to the different response process from question comprehension to response mapping.

Acknowledgements

The Saudi National Mental Health Survey (SNMHS) is carried out in conjunction with the World Health Organization World Mental Health (WMH) Survey Initiative which is supported by the National Institute of Mental Health (NIMH; R01 MH070884), the John D. and Catherine T. MacArthur Foundation, the Pfizer Foundation, the US Public Health Service (R13-MH066849, R01-MH069864, and R01 DA016558), the Fogarty International Center (FIRCA R03-TW006481), the Pan American Health Organization, Eli Lilly and Company, Ortho-McNeil Pharmaceutical, GlaxoSmithKline, and Bristol-Myers Squibb. We thank the staff of the WMH Data Collection and Data Analysis Coordination Centres for assistance with instrumentation, fieldwork, and consultation on data analysis. None of the funders had any role in the design, analysis, interpretation of results, or preparation of this paper. A complete list of all within-country and cross-national WMH publications can be found at http://www.hcp.med.harvard.edu/wmh

We also thank Beth-Ellen Pennell, Yu-chieh Lin and other staff at the Survey Research Center, University of Michigan for their advice and support on the design and implementation of the SNMHS. We are grateful to Dr. Abdullah Al-Subaie, Dr. Abdulhameed Al-Habeeb, Noha Kattan, Feda Al-Tuwaijri, the cognitive interviewers, the SNMHS staff and all the study participants. We also thank the reviewers, the editors of the special issue and Dr. Gordon Willis for their comments and suggestions, which greatly improved our manuscript.

Funding

The Saudi National Mental Health Survey (SNMHS) is under the patronage of the King Salman Center for Disability Research. It is funded by Saudi Basic Industries Corporation, King Abdulaziz City for Science and Technology, Abraaj Capital, Ministry of Health (Saudi Arabia), and King Saud University; King Faisal Specialist Hospital and Research Center, and Ministry of Economy and Planning, General Authority for Statistics are its supporting partners.

Declaration of conflicting interests

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

References

Agans, R.P., Deeb-Sossa, N., Kalsbeek, W.D. (2006). Mexican immigrants and the use of cognitive assessment techniques in questionnaire development. *Hispanic Journal of Behavioral Sciences, 28*(2), 209-230.

Andreenkova, A. (forthcoming). “Sensitive questions in comparative surveys”. In B. Dorer, T.P. Johnson, B.E. Pennell, & I. Stoop (Eds.), *Advances in comparative survey methods: Multicultural, multinational and multiregional contexts (3MC)* NJ: John Wiley & Sons, Inc.

Barnett, J. 1998. Sensitive questions and response effects: an evaluation. *Journal of Managerial Psychology, 13*(1/2), 63-76.

Beatty, P. 2004. The dynamics of cognitive interviewing. In S. Presser, J. M. Rothgeb, M. P. Couper, J. T. Lessler, E. Martin, J. Martin, & E. Singer (Eds.), *Methods for testing and evaluating survey questionnaires* (pp. 45-66). Hoboken, NJ: John Wiley.

Bernardi, R. A. 2006. Associations between Hofstede’s cultural constructs and social desirability response bias. *Journal of Business Ethics, 65*(1), 43-53.

Berrigan, D., Forsyth, B. H., Helba, C., Levin, K., Norberg, A., & Willis, G. (2010). Cognitive testing of physical activity and acculturation questions in recent and long-term Latino immigrants. *BMC Public Health*, *10*, 481: https://doi.org/10.1186/1471-2458-10-481

Blair, E. ,Sudman, S., Bradburn, M.N., & Stocking, C. (1977). How to ask questions about drinking and sex: response effects in measuring consumer behaviour. *Journal of Marketing Research*, *14*, 316-21.

Boeije, H., & Willis, G. 2013. The cognitive interviewing reporting framework (CIRF). *Methodology, 9*(3), 87-95.

Bond, R., & Smith, P. B. (1996). Culture and conformity: A meta-analysis of studies using Asch's (1952b, 1956) line judgment task. *Psychological Bulletin, 119*(1), 111.

Conrad, F., & BlairJ. (1996). From impressions to data: Increasing the objectivity of cognitive interviews. *Proceedings of the Section on Survey Research Methods, Annual Meetings of the American Statistical Association*, 1-10. Alexandria, VA: American Statistical Association.

Goerman, P. (2006). *Adapting cognitive interview techniques for use in pretesting Spanish language survey instruments*. Washington, DC: Statistical Research Division Research Report Series, Survey Methodology #2006-3, U.S. Census Bureau

Conrad, F., & Blair, J. (2001). Interpreting verbal reports in cognitive interviews: Probes matter. *Proceedings of the Annual Meeting of the American Statistical Association, Section on Survey Research Methods*. Alexandria, VA: American Statistical Association.

Harkness, J. (2011). *Translation. Cross-cultural survey guidelines*. Retrieved 12 April, 2015 from http://ccsg.isr.umich.edu/translation.cfm.

Harkness, J.A., Van de Vijver, F.J., & Mohler, P.P. (2003). *Cross-cultural survey methods* (Vol. 325). Hoboken, NJ: Wiley-Interscience.

Harkness, J.A., Villar, A., & Edwards, B. (2010). Translation, adaptation, and design. In J. A. Harkness, M. Braun, B. Edwards, T. P. Johnson, L. Lyberg, P. Ph. Mohler, … T. W. Smith (Eds.), *Survey methods in multinational, multicultural and multiregional contexts* (pp. 117-140). Hoboken, NJ: John Wiley & Sons

Heine, S. J. 2007. Culture and motivation: What motivates people to act in the ways that they do. In S. Kitayama & D. Cohen (Eds.), *Handbook of cultural psychology* (pp. 714-733). New York: Guilford Press

Johnson, T.P., & van de Vijver, F.J.R. (2003). Social desirability in cross-cultural research. In J.A. Harkness, F.J.R. van de Vijer, & P. P. Mohler (Eds.), *Cross-cultural survey methods* (pp. 195–206). Hoboken, NJ: Wiley.

Kessler, R. C., & Ustun, T. B. (2004). The world mental health (WMH) survey initiative version of the world health organization (WHO) composite international diagnostic interview (CIDI). *International Journal of Methods in Psychiatric Research 13*(2), 93-121.

Kitayama, S., Duffy, S., & Uchida, Y. (2007). Self as cultural mode of being. In S. Kitayama & D. Cohen (Eds.), *Handbook of cultural psychology* (pp. 136-174). New York: Guilford Press.

Lalwani, A. K., Shavitt, S., & Johnson, T. (2006). What is the relation between cultural orientation and socially desirable responding? *Journal of Personality and Social Psychology 90*(1), 165.

Lee, R.M. (1993). *Doing research on sensitive topics*. London: Sage.

Markus, H. R., & Kitayama, S. (1991). Culture and the self: Implications for cognition, emotion, and motivation. *Psychological Review 98*(2), 224.

Miller, K. (forthcoming). Cognitive interviewing methodology to examine survey question comparability. In B. Dorer, T.P. Johnson, B.E. Pennell, & I. Stoop (Eds.), *Advances in comparative survey methods: Multicultural, multinational and multiregional contexts (3MC)*. NJ: John Wiley & Sons, Inc.

Miller, K., Fitzgerald, R., Padilla J., Willson, S., Widdop, S., Caspar R., … Alisu, S. (2011). Design and analysis of cognitive interviews for comparative multinational testing. *Field Methods, 23*(4), 379-396. DOI: 10.1177/1525822X11414802

Potaka, L., & Cochrane, S. (2004). Developing bilingual questionnaires: Experiences from New Zealand in the development of the 2001 Mäori language survey. *Journal of Official Statistics, 20*(2), 289.

Presser, S., Couper, M. P., Lessler, J. T., Martin, E., Martin, J., Rothgeb, J. M. & Singer, E. (2004). Methods for testing and evaluating survey questions. *Public opinion quarterly*, *68*(1), 109-130.

Storti, C. (1999). *Figuring foreigners out: a practical guide*. Yarmouth, Maine: Intercultural Press, Inc.

Suzuki, N., & Yamagishi. T. (2004). An experimental study of self-effacement and self-enhancement among the Japanese. *Japanese Journal of Social Psychology, 20*(1), 17-25.

Tourangeau, R., & Yan, T. (2007). Sensitive questions in surveys. *Psychological Bulletin, 133*(5), 859-883.

Triandis, H.C. (1995). *Individualism and collectivism*. Boulder, CO: Westview Press.

Uskul, A. K., Oyserman, D., & Schwarz, N. (2010). Cultural emphasis on honor, modesty, or self-enhancement: Implications for the survey response process. In J. A. Harkness, M. Braun, B. Edwards, T. P. Johnson, L. Lyberg, P. Mohler, … T. W. Smith (Eds.), *Survey methods in multinational, multiregional, and multicultural contexts* (pp. 191-202). Hoboken, NJ: John Wiley & Sons, Inc.

Vernon, M. K. (2005). *Pre-testing sensitive questions: Perceived sensitivity, comprehension, and order effects of questions about income and weight*. Washington, DC: Bureau of Labor Statistics.

Warnecke, R.B., Johnson, T.P., Sudman, S., O'Rourke, D.P., Lacey, L., & Horm, J. (1997). Improving question wording in surveys of culturally diverse populations. *Annals of Epidemiology, 7*(5), 334-42.

Willis, G. B. (1997). The use of the psychological laboratory to study sensitive survey topics. In L. Harrison & A. Hughes (Eds.), *The validity of self-reported drug use: Improving the accuracy of survey estimates* (pp. 416-438). USA: National Institutes of Health.

Willis, G. B. (2005). *Cognitive interviewing: A tool for improving questionnaire design*. Thousand Oaks, CA: Sage.

Willis, G.B. (2015a). The practice of cross-cultural cognitive interviewing. *Public opinion quarterly*, *79*(Special Issue), 359-395.

Willis, G. B. (2015b). *Analysis of the cognitive interview in questionnaire design.* NY: Oxford University Press.

Willis, G. B., & Miller, K. (2011). Cross-Cultural Cognitive Interviewing: Seeking Comparability and Enhancing Understanding. *Field Methods, 23*(4), 331-341. DOI: 10.1177/1525822X11416092

Willis, G. B., Kudela, M. S., Levin, K. L., Norberg, A., Stark, D. S., Forsyth, B. H., … Hartmann, A.M. (2010). Evaluation of a multistep survey translation process. In J. A. Harkness., M. Braun, B. Edwards, T. P. Johnson, L. Lyberg, P. Ph. Mohler, … T. W. Smith (Eds.), *Survey methods in multinational, multicultural and multiregional contexts* (pp. 137-152). Hoboken, NJ: John Wiley & Sons.

Willis, G., Lawrence, D., Hartman, A., Kudela, M. S., Levin, K., & Forsyth B. (2008). Translation of a tobacco survey into Spanish and Asian languages: The tobacco use supplement to the current population survey. *Nicotine & Tobacco Research*, *10*(6), 1075–1084. https://doi.org/10.1080/14622200802087572

Yamaguchi, S. (1994). Collectivism among the Japanese: A perspective from the self. In U. Kim, H. C. Triandis, C. Kagitcibasi, S-C. Choi, & G. Yoon (Eds.), *Individualism and collectivism: Theory, method, and applications* (pp. 175-188). London: Sage.

Appendix:

The table below summarises the sample questions probed for sensitivity in Cognitive Interviews. The questions are grouped by section depending on how they appear in the questionnaire. The Arabic translation is provided facing the original English wording.

Table A.1: Sample questions probed for sensitivity in Cognitive Interviews.

Original English	Adapted Arabic
30 Day Symptoms Section	
NSD7g. How often did you feel worthless?	كم كنت عادة تشعر بأن ليس لك قيمة؟
NSD13c. How often were you so angry that you felt out of control?	كم كنت عادة تغضب لدرجة أنك شعرت أنك خرجت عن السيطرة؟
Depression Section	
D26aa. Did you often think about death, either your own, someone else's, or death in general?	هل كنت تفكر بالموت، سواء موتك أنت أو موت أي شخصِ آخر أو الموت بشكل عام؟
D84c. How many professionals did you ever talk to about your (low mood/**key phrase**)?	كم عدد المختصين على الإطلاق اللذين تحدثت معهم عن حالة (المزاج المنخفض/**العبارة الرئيسية**)؟
Mania Section	
M7b. Were you a lot more interested in sex than usual, or did you want to have sexual encounters with people you wouldn't ordinarily be interested in?	هل كان لديك اهتمام زائد عن المعتاد بالجنس أو هل كنت ترغب في ممارسات جنسية مع أشخاص ليس لك فيهم رغبة في الأحوال العادية؟
M7o. Did you have the idea that you were actually someone else, or that you had a special connection with a famous person that you really didn't have?	هل كنت تعتقد أنك شخص آخر فعلاً، أو أن لك علاقة خاصة بشخص مشهور لا تربطك به أي علاقة في الحقيقة؟
Panic Disorder Section	
PD1n. Were you afraid that you might die?	هل كنت تخاف أنك قد تموت؟
Specific Phobia	
SP9f. Did you ever fear that you might lose control, go crazy, or pass out?	هل كنت تخاف أن تفقد السيطرة على نفسك أو تفقد عقلك أو يغمى عليك؟
Social Phobia	
SO6.1. Was there ever a time in your life when you felt emotionally upset, worried, or disappointed with yourself because of your fear (or avoidance) of (this situation/ these situations)?	هل سبق أن مر عليك فترة في حياتك شعرت فيها بالانزعاج النفسي أو القلق أو الإحباط من نفسك بسبب رهبتك أو تجنبك لـ(هذا الموقف/ هذه المواقف)؟
Suicidality Section	
SD2. Three experiences are listed in your booklet on page 20 labeled A, B, and C. Did experience A ever happen to you? EXPERIENCE A IS 'YOU SERIOUSLY THOUGHT ABOUT COMMITTING SUICIDE'	تجد في صفحة 20 من كتيبك ثلاث تجارب مرقمة بـ أ و **ب** و **ج**. هل سبق أن مرت بك التجربة أ ؟ **التجربة أ هي "فكرتُ جدياً بالانتحار"**
SD3. Three experiences are listed in your booklet on page 20 labeled A, B, and C. Did Experience A happen to you at any time in the past 12 months? EXPERIENCE A IS 'YOU SERIOUSLY THOUGHT ABOUT COMMITTING SUICIDE'	تجد في صفحة 20 من كتيبك ثلاث تجارب مرقمة بـ أ و **ب** و **ج**. هل مرت بك التجربة أ في أي وقت خلال الاثنى عشر شهراً الأخيرة؟ **التجربة أ هي "فكرتُ جدياً بالانتحار"**

Table A.1 (continued)

Original English	Adapted Arabic
Premenstrual Syndrome Section	
PR8. Have you ever taken hormone replacement therapy pills for symptoms of menopause?	هل سبق أن أخذتِ حبوب الهرمونات التعويضية لأعراض انقطاع الدورة الشهرية؟
Illegal Substance Use Section	
IU1b. Was your use ever so regular that you felt you could not stop using the sedative or tranquilizer prescribed for you?	هل كنت تستخدم المهدئات بشكل مستمر لدرجة أنك شعرت أنك لا تستطيع التوقف عنها؟
IU25. Were you arrested or stopped by the police more than once because of driving under the influence of [DRUG] or because of your behavior while you were under the influence of [DRUG]?	هل سبق أن قبضت عليك الشرطة أو المرور أو الدوريات أو أوقفوك أكثر من مرة بسبب قيادة السيارة تحت تأثير (**العقار**) أو بسبب تصرفاتك وأنت تحت تأثير (**العقار**)؟
Conduct Disorder Section	
CD16c. How often did you use a weapon on another person, like a baseball bat, glass bottle, knife, gun, or brick?	كم كنت عادة تستخدم السلاح ضد أحد مثل العصا أو قارورة أو سكين أو مسدس أو حصاة؟
CD16f. How often did you force someone to give you something like money, jewellery, or clothing by threatening them or causing them injury?	كم كنت عادة تجبر أحد بواسطة التهديد أو الإيذاء الجسدي على أن يعطيك شيء ما مثل المال أو المجوهرات أو الملابس؟