Published in final edited form as:

Field methods. 2011 ; 23(4): 439–460. doi:10.1177/1525822X11418176.

Using Cognitive Interviewing and Behavioral Coding to Determine Measurement Equivalence across Linguistic and Cultural Groups: An Example from the International Tobacco Control Policy Evaluation Project

James Thrasher, PhD, MA, MS[1,2], **Anne C.K. Quah, PhD**[3], **Gregory Dominick, MA**[1], **Ron Borland, PhD**[4], **Pete Driezen, MS**[5], **Rahmat Awang, PharmD**[6], **Maizurah Omar, PhD**[6], **Warwick Hosking, PhD**[7], **Buppha Sirirassamee, PhD**[8], **Marcelo Boado, PhD**[9], and **Kristen Miller, PhD**[10]

[1]Department of Health Promotion, University of South Carolina, USA [2]Instituto Nacional de Salud Pública, México [3]Department of Psychology, University of Waterloo, Canada [4]Cancer Council of Victoria, Melbourne, Australia [5]Population Health Research, University of Waterloo, Canada [6]National Poison Centre, Universiti Sains Malaysia, Malaysia [7]School of Psychology, Victoria University, Melbourne, Australia [8]Institute for Population and Social Research, Mahidol University, Thailand [9]Departamento de Sociología, Universidad de la República, Montevideo, Uruguay [10]National Center for Health Statistics, Centers for Disease Control, USA

Abstract

The present study aimed to examine and compare results from two questionnaire pretesting methods (i.e., behavioral coding and cognitive interviewing) in order to assess systematic measurement bias in survey questions for adult smokers across six countries (USA, Australia, Uruguay, Mexico, Malaysia and Thailand). Protocol development and translation involved multiple bilingual partners in each linguistic/cultural group. The study was conducted with convenience samples of 20 adult smokers in each country. Behavioral coding and cognitive interviewing methods produced similar conclusions regarding measurement bias for some questions; however, cognitive interviewing was more likely to identify potential response errors than behavioral coding. Coordinated survey qualitative pretesting (or post-survey evaluation) is feasible across cultural groups, and can provide important information on comprehension and comparability. Cognitive interviewing appears a more robust technique than behavioral coding, although combinations of the two might be even better.

Introduction

Cross-cultural research often assumes that measurement is equivalent across cultural groups (Bollen, Entwisle, & Alderson, 1993; T. W. Smith, 2004). Assurance that construct meaning is equivalent across groups is a fundamental departure point for this assumption (Johnson, 1998; Van de Vijver, 2004; Van de Vijver & Leung, 1997). When constructs are measured with multiple questions, factor analytic techniques may be used to establish the equality of construct parameters (e.g., dimensionality, factor loadings, scaling) across groups and, thereby, validate group comparisons (Burlew, Feaster, Brecht, & Hubbard, 2009; Gregorich, 2006; Van de Vijver & Leung, 1997; Van Herk, Poortinga, & Verhallen, 2005). However, these techniques cannot be used when constructs are measured with a single question, as in survey research where multi-item scales are often not practical. For constructs measured with a single survey question, special attention should be paid to how biases in measurement may vary systematically across cultural groups (T. W. Smith, 2004).

Systematic measurement bias, which we define as the bias component of measurement error that varies differentially across cultural groups, may occur at the level of the construct, method, and item (Van de Vijver & Leung, 1997). As previously stated, the meaning of the construct must be equivalent across cultural groups. Even so, any single question used to measure that construct may be more or less relevant to construct meaning in one cultural group compared to another, thereby biasing measurement. Further, the method of data collection may bias measurement if social desirability, response styles, stimulus familiarity, or interviewer effects vary across cultural groups. Finally, at the level of the item or question, poor translation or inadequate item formulation (e.g., complex wording) may introduce systematic measurement bias, as might varying ranges of connotations of meaning for terms used. The validity of cross-cultural comparison depends on minimizing these biases.

Coordinated translation and pretesting of questionnaires may help identify and thereby reduce the likelihood of systematic measurement bias (Goerman, 2006; Harkness, Pennell, & Schoua-Glusberg, 2004; Miller, Mont, Maitland, Altman, & Madans, in press; Thrasher & Johnson, 2008; Willis & Zahnd, 2007). Question pretesting methods can also be used in post-hoc fashion to inform interpretation of results and to assist with future question improvement (Willis et al., 2008). Results from these methods have been used to identify systematic measurement biases due to translation, cultural adaptation, and more general questionnaire design issues (Willis et al., 2008; Willis & Zahnd, 2007). This and the Van de Vivjer and Leung schemas are useful, but neither is definitive in providing an interpretive framework for cross-cultural measurement research.

a. Cognitive interviewing.

Cognitive interviewing is increasingly used to assess and remediate problems with survey questions that contribute to response error, which we define as error in measurement, whether random (which leads to less precise estimates) or systematic (which biases estimates). By prompting study participants to provide information about the response process, cognitive interviewing aims to identify breakdowns that contribute to response error (Willis, 2005). However, the scientific literature is mainly suggestive concerning how to apply this methodology across cultural contexts (Willis et al., 2008). Furthermore, some probing techniques that are used identify breakdowns in the survey response process may work better in some cultural groups than others (Goerman, 2006). Nevertheless, emerging research suggests that careful cultural adaptation can help overcome some of these barriers (Goerman, 2006; Miller et al., in press; Pan, Craig, & Scollon, 2005; Willis et al., 2008; Willis & Zahnd, 2007).

b. Behavior coding.

Behavior coding was originally developed to evaluate interviewer performance (Cannell CF, 1975) but is increasingly used to assess response error (Fowler, 1995; Johnson et al., 2006). Respondent behaviors are evaluated either during the interview or afterwards using recordings (Esposito, Rothgeb, Polivka, Hess, & Campanelli, 1992; Hughes, 2004; Miller et al., in press). Unlike cognitive interviewing, behavior coding does not rely on probing. Rather, behavior codes that register participant difficulties include asking for questions to be repeated, providing a response that is not included amongst response options, and asking for clarification of the question (Miller et al., in press; Van der Zouwen & Smit, 2004). Researchers ascertaining systematic measurement error across groups have used behavioral coding by itself (Johnson et al., 2006), as a subsequent phase after conducting cognitive interviews and changing questions (Willis et al., 2008), and simultaneously with scripted cognitive interviewing probes (Miller et al., 2009).

Behavior coding provides systematic, objective information that can be compared quantitatively across cultural groups. Nevertheless, the meaning of overt behaviors may vary considerably across cultural contexts. Furthermore, particular respondent behaviors (e.g., request for clarification) may not be associated with response error. Given these provisos, the triangulation of behavioral coding results with those obtained from qualitative pretesting methods could help in the interpretation of results and provide more information on which to base conclusions about response error and its source (Forsyth, Rothgeb, & Willis, 2004; Zahnd et al., 2005).

c. The study context:

The International Tobacco Control Policy Evaluation Project (ITC Project), aims to evaluate the impact of World Health Organization-Framework Convention on Tobacco Control policies (WHO, 2003) among adult smokers (Fong et al., 2006; Thrasher et al., 2006). Using a quasi-experimental research design, the ITC Project compares behavioral and psychosocial data from cohorts of smokers in countries with and without particular policies of interest (Thompson et al., 2006; Thrasher, Boado, Sebrié, & Bianco, 2009). ITC survey questions have relatively well-established validity in the four Anglo countries where the ITC Project began (IARC, 2009; Thompson et al., 2006).

As the ITC Project has expanded to other countries, the survey has been translated into 14 additional languages, as well as modified to be consistent with various language varieties (e.g., Uruguayan Spanish), the latter primarily relying on the language competencies of the within-country investigators. Countries have gradually, and often hurriedly, joined the Project, usually because of contingencies around upcoming legislation. Although, simultaneous, coordinated translation was not possible, best-practices committee translation methods (Harkness, 2003; Pan & de la Puente, 2005) were followed in some countries (e.g., Mexican Spanish). Some countries used professional translators (e.g., Malaysia), with bilingual investigators reviewing and adjusting to capture intended meaning, whereas in other countries (e.g., Thailand), bilingual investigators who were experts in survey content translated the survey in other countries. Data reported here were collected to assess and improve ITC survey questions for which investigators reported evidence of response error due to any of sources of bias described above. Herein, we report on the feasibility of these pre-testing methods and interpret the results they produce. Hypotheses regarding the type and distribution of response error across countries depended on the question. In general, however, we hypothesized that acquiescent individuals would have less of a tendency to overtly indicate where they have problems with a survey question. That is, they would be less likely to “challenge” the interviewer by objecting that a question may have some defect through requests for clarification or by displaying other behavioral cues. Given higher acquiescence, or “yea saying,” found in Asian and Latino cultures than in Anglo cultures (P. Smith, 2004), we expected a lower frequency of problems indicated by behavioral codes in Malaysia, Thailand, Mexico and Uruguay than in the US and Australia.

Methods

Protocol development:

Research teams involved in the original translation and data collection for the ITC Project in the US, Australia, Mexico, Uruguay, Thailand and Malaysia participated in this study. In developing the protocol, we followed the steps outlined in Figure 1. To begin, investigators in each country were asked to identify questions for which they had any evidence of problems from field interviewers, data analysts or people involved in translation, as well as their rationale for selecting these questions. A list of 26 potentially problematic questions was established, for which we developed follow-up questions or “scripted probes” to target specific issues of interpretation (i.e., construct or item bias), evaluation (i.e., method bias) and decision-making (i.e., method and item bias). Some of these probes allowed for open responses (e.g., *What does it mean to say that something is “addictive”?*), whereas others involved fixed-choice responses (e.g., *Thinking about the last 6 months, do you find it difficult to remember if you saw, or did not see, any tobacco advertising on television?*). As in some other cross-cultural studies (Miller et al., 2009), scripted probes were chosen over emergent probes (i.e., the interviewer uses a flexible array of probes, depending on participant cues) or verbal reports (i.e., asking participants to verbalize their thoughts when responding to the question). Some people are better at verbal reports than others, and Anglo participants may be better than people from other cultural groups (Coronado & Earle, 2002). Furthermore, only one interviewer had experience with cognitive interviewing, and experience is critical to the sound use of emergent probes (Willis, 2005). The use of scripted probes ensured a relatively standardized application of the protocol. The final data collection instrument consisted of: 1) the subset of 26 original ITC questions; 2) interviewer-determined behavioral codes (i.e., participant needed the question repeated; had difficulty with response options; asked for clarification or qualified response) following each original ITC question; and 3) question-specific scripted cognitive probes. Data on each of these three domains were gathered from every participant.

Once the draft English-language protocol was ready, we developed a “translator’s guide” to provide a definition of the concept that original ITC survey questions presumably measured, the rationale for each scripted probe, and indications of the country-specific nature of the problem. The original ITC questions were included in the protocol exactly as phrased in the corresponding ITC survey instrument for each country. Bilingual investigators in each country then provided comments on the draft protocol and translator's guide. Protocols were then translated and adjusted to address additional issues that arose during its translation.

Training with country project coordinators and cognitive interviewers took place through a conference call that was guided by an accompanying PowerPoint presentation and video spots that modeled correct and incorrect interview administration. Cognitive interviewers in each country then piloted the protocol with two participants each, audio recording the interview and entering quantitative and qualitative data into an Excel spreadsheet. The project coordinator reviewed the audio and resulting data, communicating concerns and needs to adjust practices to the relevant country coordinators and interviewers. Once we determined that interviewers could adequately administer the protocol, we received Internal Review Board approval.

Sample:

Recruitment of study participants took place in one city in each country (Columbia, South Carolina, USA; Melbourne, Australia; Cuernavaca, Mexico; Montevideo, Uruguay; and Penang, Malaysia, respectively) except Thailand, where participants were recruited from four rural provinces, as well as Bangkok, because of investigator concerns that rural respondents had experienced particular response difficulties. Street intercept methods were used to recruit convenience samples of 20 adults smokers in each country, whereby staff asked passers by of their interest in participating in a study on smoking and, if interested, their eligibility (i.e., 18 or older, smoked at least once in previous month, smoked at least 100 lifetime cigarettes). As is standard practice before beginning cognitive interviews (Willis, 2005), participants were briefed regarding the goal of determining problems with the questions, not with participants' responses to the questions. Recruitment took place near a quiet locale, such as a library or café, where interviews were conducted. Interviews were audio recorded and transcribed to capture open-ended responses and to verify data entry. Participants were compensated in different ways, depending on country norms (i.e., $20 cash in the US and Australia; $10 cash in Malaysia and Thailand; phone cards worth approximately $10 USD in Mexico and Uruguay).

The percentage of male participants generally reflected the gender distribution of smoking in each country, with more male participants in Malaysia and Thailand (see Table 1). The age of participants ranged from 18 to 75, with mean ages for each country ranging from 31 in Malaysia to 40 in Uruguay. Levels of education were generally comparable, except for lower educational achievement among Thai participants. The percentage of daily smokers in each country was generally equivalent (i.e., 80% to 90%), as was the mean number of cigarettes smoked each day (13–15/day), except for higher frequency of smoking in the US sample (23/day).

Analyses

Analysis of responses to the open-ended cognitive probes involved standard content analysis techniques for determining dominant themes of narrative segments (Miles & Huberman, 1994). Audio-recorded responses were transcribed in the original language and then translated into English. Research assistants and two PIs (JFT and ACKQ) examined English transcripts to inductively determine dominant themes, which were defined and given a specific code (see Tables 3 and 4). Each response was then reviewed and assigned all codes that captured its content. Teams in non-English speaking countries then reviewed results to determine whether these codes adequately registered the content of the original language transcription. If there were discrepancies, teams worked with the coordinating center to reach consensus on which codes and code definitions best represented the original-language transcription. Once consensus was reached on the codes and their appearance within transcripts, we examined the frequency of codes associated with each open-ended response, by country.

Due to the small sample sizes within each country, Fisher's exact tests were used to assess country-level differences in behavioral codes, responses to fixed-choice cognitive probes, and the frequency of themes that occurred in response to open-ended cognitive probes. Because the study was not powered for such tests, we generally interpret statistically significant results as indicating a relatively large difference between countries. When the results indicated significant variation across groups, we examined the patterns within the data to determine which countries, if any, were most similar to and distinct from one another with regard to these indicators.

Results:

Selection of questions for examination:

Of the 26 ITC survey questions assessed, behavioral codes indicated potential systematic measurement bias for 46% (12/26). Analysis of the open-ended and fixed-choice cognitive probes indicated evidence of systematic measurement bias for 56% (9/16) and 58% (10/17) of the questions, respectively, with one or the other probe type indicating possible systematic bias for 72% (18/26) of questions where they were used. In the questions for which both cognitive probes and behavioral coding were used (n=25), the results coincided in indicating systematic measurement bias for 36% of the questions and its absence for 20%.

For 17 questions, behavioral codes and fixed-choice probes were used, and for 16 questions behavioral codes and open-ended probes were used. For purposes of the current study, we selected questions that represented convergence and divergence of results regarding systematic measurement bias at the level of the question (not the level of the individual response to the question). Where convergence was of interest for fixed-choice (12/17) and open-ended (7/16) probes, we selected representative examples from among those that indicated problems. When focusing on divergent results, we selected a question where behavioral coding was less likely to register systematic measurement bias than cognitive probing, which was the more common pattern of divergence.

a. Agreement between behavioral coding and fixed-choice probing—As an indicator of descriptive social norms, the prevalence of smoking among proximal social network members was assessed with the question "*Of the five closest friends or acquaintances that you spend time with on a regular basis, how many of them are smokers?"* with responses from 0 to 5. We anticipated potential issues with the double-barrelled nature of the question ("friends or acquaintances"), the potential vagueness of the referent "acquaintance," the difficulty of limiting the universe of possible referents to the five "closest friends or acquaintances," and the frequency as expressed as a "regular basis." Fixed-choice probes aimed to assess these issues, including the additional linguistic and cultural issue of whether some groups were more or less likely to consider family members in responding to this question.

Behavioral codes (Table 2) indicated relatively equivalent frequencies of participant requests to repeat the question, ask for clarification, or to qualify their answer. However, participants had greater difficulty using the response options in the US (n=8/20) and Thailand (n=6/20) than in the other countries (p=0.001). Participants in all countries said that the question was difficult to respond to (Australia n=5; US n=4; Uruguay n=5; Mexico n=2; Malaysia n=3; Thailand n=2), although the frequency of this issue was not statistically different across countries (P=0.72). When participants were asked why it was difficult, they primarily identified three reasons: 1) Difficulty in deciding the top five friends (n=8); 2) Difficulty knowing the smoking status of people described, including whether friends they had not seen in a while had quit or not (n=6); 3) Difficulty deciding on the acquaintances you spend the most time with (n=2). Assessment of the frequency of these reasons by country found no statistically significant differences, although the sample size for this participant subset was particularly small.

The next cognitive probe aimed to address the frequency of contact with the referents (i.e., *Can you tell me how often you see each of the five people you thought about when answering this question? daily; at least once a week; at least once a month; at least once a year; less often*). Overall, 77% of the people that respondents focused on were people who they saw at least once a week. The results for this characteristic of the question indicated no strong evidence for systematic measurement bias across countries.

Finally, we were interested in determining whether the connotation of the terms used to capture "friends and acquaintances" encompassed family members or not. The Fisher's exact test indicated a statistically significant difference in responses (p<0.001), with the clearest difference for the Thai participants, none of whom considered family among their friends. It was determined that the Thai term used to capture "friend" connotes only non-kin, as it involves a linguistic term that encodes relative social standing that excludes family. Otherwise, about half of the Australian, US and Malaysian respondents (n=11, 11, and 10, respectively) and three quarters of the respondents from Mexico and Uruguay (n=15 and 14, respectively) stated that family could be considered among friends and acquaintances.

Summary: Both behavioral coding and scripted probes indicated problems with comprehension. Behavior codes that focused on issues with the response options were more frequent in the US and Thailand than in other countries, whereas the frequency of other problems indicated by behavioral codes were similar. Fixed-choice cognitive probes also indicated decision-making issues across all countries, particularly with respect to limiting the universe of appropriate social referents. Evidence for systematic measurement bias across countries was most salient when assessing whether family members were potential referents for the question. Responses were similar across countries, except in the cast of Thailand, where the term connoted non-kin. Such issues with differing connotations of linguistic terms are unlikely to be registered with behavioral codes, and are more related to translation and construct definition than to basic question design.

b. Agreement between behavioral coding and open-ended probing—

Knowledge of the dangers associated with the depth of inhaling cigarette smoke was assessed with the question: *Is the following statement true or false? The way a smoker puffs on a cigarette can affect the amount of tar and nicotine a smoker takes in*. We developed two probes to assess understanding of "the way a smoker puffs" (i.e., depth of cigarette smoke inhalation) and "smoker takes in" (i.e., absorb cigarette smoke chemicals into lungs and blood stream). Responses to the original ITC question indicated that participants in the US, Uruguay and Mexico were less likely than participants in Australia, Malaysia and Thailand to view the response as true (P=0.001). The frequency of behavioral codes regarding participant problems with response options and asking for clarification or qualifying the answer was higher for US and Australian participants than those in other countries (Table 2).

Through cognitive probing, participants were asked to explain the different "ways that a smoker puffs" on a cigarette, using the same phrasing as in the original ITC question. Responses were examined for evidence of misunderstanding the concept of depth of inhalation. Results indicated a marginally significant difference across countries (P=0.053), with lower levels of intended understanding of the concept in Uruguay (n=12) than in other countries (Australia=19; US=17; Mexico=17; Malaysia=16; Thailand=19).

We also asked participants to tell us what they know about what happens to cigarette smoke that a smoker "takes in", using the same phrasing as in the original ITC question. Overall we found relatively low levels of intended understanding of the underlying concept. More participants understood this concept in the US, Australia and Mexico (n=16, 19 and 15, respectively) than in Malaysia, Uruguay and Thailand (n=11, 10 and 5, respectively).

Summary: Both behavioral codes and the open-ended probing indicated systematically different problems across countries. However, the results diverged when specific countries where problems occurred were examined. Behavioral coding results indicated more evidence of problems in the US and Australia, whereas the cognitive probes indicated problems for Uruguay in understanding either of "ways of puffing" or "taking in", and the terms used to capture "taking in" also appeared problematic for Malaysian and Thai participants. Discussion of this issue with bilingual team members in these countries suggested the difficulty of translating these concepts in simple terms. Although the Mexican Spanish terms used to translate the concept "take in" appeared relatively successful, it had been changed due to concerns that Uruguayan Spanish speakers would not understand it. Nevertheless, the adaptation does not appear to have been adequate.

c. Disagreement between behavioral coding and open-ended cognitive probing—

Perceptions of the addictiveness of tobacco were assessed with the item "*Tobacco is addictive,"* with a 5-point Likert scale indicating extent of agreement. We were concerned with the technical and vague nature of the term "addiction." Overall, 90% of all participants agreed or strongly agreed with the statement, with some disagreement found in the US, Uruguay, Mexico and Malaysia.

Behavioral codes were indicated only a few times, with no indication of systematic differences in their frequency across countries. When participants were asked, through cognitive probing, to clarify the meaning of the term “addiction”, the following themes best characterized the content of responses:

- General control: general ability to restrain impulses to smoke
- Physiological control: ability to restrain impulses to smoke, with a focus on biological or bodily impulses
- Psychological control: ability to restrain impulses to smoke, with a focus on mental processes
- Control over quitting: ability to quit when desired
- Frequency/quantity: amount or regularity of smoking
- Danger: negative outcomes of smoking
- Pleasure: pleasurable nature of smoking

The frequency with which these themes occurred in participants descriptions of addiction were analyzed (Table 3). General control was a primary theme across all countries, with no significant variation across countries. However, a specific focus on physiological issues appeared more prevalent in Australia and the US and a focus on psychological dimensions was more prevalent in Malaysia, Australia and US. A focus on the frequency or quantity of consumption was also a dominant theme that was found more frequently in Australia, the US, and Thailand. Finally, a focus on danger was prevalent in Thailand (n=7), but not in the other countries.

To further examine the semantic domain of addiction, we asked participants to tell us about other things, besides tobacco, that are addictive (Table 4). Alcohol was salient in Australia, the US, Uruguay and Mexico, but less so in Malaysia and Thailand. Illicit drugs were also frequently mentioned in all countries except Malaysia. Furthermore, pleasurable foods were mentioned with more frequency in the US and Uruguay than in other countries.

Summary: Behavioral coding did not indicate either general response error issues or systematically different measurement bias in participants reactions to this question. However, open-ended responses to cognitive probes that aimed to uncover the meanings of and associations with the term “addiction” suggested differential understanding of the term across groups, with the semantic domain of addiction encompassing different factors in different countries. In the end, cross-cultural study of complex concepts like addiction may be particularly susceptible to bias around the equivalence of the construct. If measured, multiple questions should probably be used to adequately register the range of behaviors and meanings that the construct connotes, while following general questionnaire design principles around simple phrasing that avoids jargon and vague terminology.

Discussion

The present study indicates that the use of behavioral coding and cognitive interviewing to assess response error and systematic measurement bias across cultural, national, and linguistic groups is logistically complicated but can be accomplished. Although cognitive interviewing methods have provoked reactance and irritation among some non-English speaking participants (Coronado & Earle, 2002; Kissam, Herrera, & Nakamoto, 1993), we did not experience this within the diverse sociocultural settings where the study was conducted. This potential issue may have been addressed by stating our goals to assess the quality of our questions, not the veracity of participants' responses. Similar approaches have facilitated participant cooperation in other cognitive interview studies on translated questionnaires (Goerman, 2006).

Overall, behavioral indicators of problems were most frequently found among US and Australian participants. This result supports the hypothesis that culturally patterned acquiescence to the perceived demands of the interview context (Knowles & Condon, 1999) may reduce the likelihood of displaying overt behaviors that register survey comprehension problems, making behavioral coding less useful outside of non-Western settings. However, behavioral codes indicating potential response error turned up in all six countries, and for a few questions were more frequent in non-Western than Western countries. When viewing convergence of behavioral coding and cognitive interviewing results as a validation check, our results indicate that behavioral codes appear to register some meaningful systematic variation across all countries; however, in one example where behavioral coding indicated systematic differences, follow-up cognitive probes indicated that comprehension problems resided in countries where behavioral coding did not register problems. Because we did not delve deeper into this issue, it is difficult to say whether behavioral coding was a less sensitive indicator than cognitive interviewing or whether it registered comprehension or response issues that our scripted probes did not address.

Overall, scripted probing indicated more problems with systematic measurement bias than behavioral coding. As with other comparative assessments of methods to detect problems with questions (Forsyth et al., 2004), it is difficult to determine the validity of these problems, particularly for measures of psychosocial constructs. Cognitive interviewing, which is explicitly designed to expose breakdowns in the question comprehension and response process, may be more likely to indicate problems when there really are none (i.e., false positives). Behavioral coding, as a more passive method of problem detection, may be more likely to provide false negative results due to its not picking up "silent misinterpretation" of the question for which there are no behavioral correlates (DeMaio & Rothgeb, 1996). This appears to have been the case, for example, with the question that focused on addiction. As suggested in the preceding paragraph, silent misinterpretation may be more prevalent in cultural settings where codes of social interaction constrain behavioral expressions of response difficulties that coding regimes developed in the English-speaking West were designed to pick up.

The potential for cognitive interviewing to address the issue of silent misinterpretation suggests that efforts should continue to refine and adapt cognitive interviewing methods to better “fit” with participants’ expectations and conversational styles (Goerman, 2006). Moreover, the results from cognitive interviewing provide more useful guidance than results from behavioral coding when determining a source of response error and suggesting ways to resolve it, although when participants exhibit behaviors, they could be actively queried to clarify why they did so (Blair, Ackerman, Piccinino, & Levenstein, 2007). Convergent results regarding the presence of bias may not always be expected, as was illustrated in our examples of differing connotations of linguistic terms that did not cause respondents to exhibit behavioral codes. Further, the process of cognitive interviewing may actually promote behaviors registered by behavioral codes that are not exhibited in the normal interview context. Nevertheless, in some instances, the triangulation of results from behavioral coding and cognitive interviewing may help tip the balance of evidence in favor of or against response error, whether general or systematic

Our results are limited by a number of factors, including the use of relatively inexperienced interviewers. Hence, we used a scripted protocol that anticipated specific response errors and may have missed other problems. Ideally, experienced interviewers would probe participants’ responses in a more flexible manner, using behavioral cues as well as their responses to questions to probe further where uncertainty still existed. However, few people are trained in these methods outside of Western contexts, so this more sophisticated approach may not always be practical.

The small convenience samples in this study are typical for cognitive testing, but do limit the generalizability of results. In all countries but Thailand, adult smokers were recruited from a single city. The Thai sample was perhaps more representative of the general population, but Thai participants were less educated than participants in other countries, who generally shared sociodemographic profiles. This difference may have biased results through factors like differential social desirability or interview context effects that reflect the lack of familiarity with survey response (Van de Vijver & Leung, 1997). Although the Malaysian and Thai participants were mostly male, this reflects the prevalence of smoking in these countries. Stratified analyses of more homogeneous samples across countries may have increased the internal validity of our assessment; however reaching adequate sample size within diverse population strata would have been costly. A final limitation is that the survey questions are inevitably asked out of context in such work, as the probing changes the context and is likely to reduce the contextualization effects of previous questions form the actual survey. In this study where we selected a sub-sample of questions for study, survey context was even further disturbed.

Overall, our findings suggest that protocols like ours are feasible and may help assess measurement bias in cross-cultural survey research. This approach may best assess bias when translation follows best-practices, well-trained interviewers are used, and sample sizes allow for subgroup analyses. Nevertheless, inadequate resources and time may mitigate against this. Our findings suggest that behavioral codes can register problems across cultural settings, although further investigation should examine the cross-cultural comparability of behavioral codes. Further research should also determine possibilities for cultural tailoring of cognitive interviewing strategies to best suit the social and conversational norms surrounding social interactions. This research should aim to assess the validity and equivalence of results when different protocols are tailored to address culturally-specific issues that may be manifested in one country but not in another.

Assessment of survey questions of the kind described should ideally take place before surveys are deployed. However, they can also play a vey useful role in assessing possible problems with survey items which are identified either in the fieldwork or in the analyses. Because post-implementation work can focus on areas of potential problems, it may be more practical in many cases.

References:

Blair J, Ackerman A, Piccinino L, & Levenstein R (2007). Using behavior coding to validate cognitive interview findings.Paper presented at the Proceedings of the American Statistical Association: Section on Survey Research Methods.

Bollen KA, Entwisle B, & Alderson AS (1993). Macrocomparative research methods. Annual Review of Sociology, 19, 321–351.

Burlew AK, Feaster D, Brecht M-L, & Hubbard R (2009). Measurement and data analysis in research addressing health disparities in substance abuse. Journal of Substance Abuse Treatment, 36, 25–43. [PubMed: 18550320]

Cannell CF, L. S, Hausser DL. (1975). A Technique for Evaluating Interviewer Performance. Ann Arbor, MI: University of Michigan.

Coronado I, & Earle D (2002). Effectiveness of the American Community Survey of the U.S. Census in a borderlands colonia setting. Washington, DC: US Census Bureau.

DeMaio TJ, & Rothgeb JM (1996). Cognitive interviewing techniques: In the lab and in the field In Schwarz N & Sudman S (Eds.), Answering questions (pp. 177–195). San Francisco, CA: Jossey-Bass.

Esposito JL, Rothgeb J, Polivka AE, Hess J, & Campanelli PC (1992). Methodologies for evaluating survey questions: Some lessons from the redesign of the Current Population Survey. Paper presented at the International Conference on Social Science Methodology, Trent.

Fong GT, Cummings KM, Borland R, Hastings G, Hyland A, Giovino GA, et al. (2006). The conceptual framework of the International Tobacco Control Policy Evaluation Project. Tobacco Control, 15(Supp 3), iii3–iii11. [PubMed: 16754944]

Forsyth B, Rothgeb JM, & Willis GB (2004). Does pretesting make a difference? An experimental test In Pressler S, Rothgeb JM, Couper MP, Lessler JT, Martin E, Martin J & Singer E (Eds.), Methods for Testing and Evaluating Survey Questionnaires (pp. 525–546). Hoboken, NJ: Wiley & Sons, Inc.

Fowler F (1995). Improving Survey Questions: Design and Evaluation. Thousand Oaks, CA: Sage.

Goerman P (2006). Adapting cognitive interview techniques for use in pretesting Spanish language survey instruments (No. Survey Methodology #2006–3). Washington, DC: US Census Bureau.

Gregorich SE (2006). Do self-report instruments allow meaningful comparisons across diverse population groups? Testing measurement invariance using the Confirmatory Factor Analysis framework. Medical Care, 44(Supp 3), S78–S94. [PubMed: 17060839]

Harkness JA (2003). Questionnaire translation In Harkness JA, Van de Vijver FJR & Mohler PP (Eds.), Cross-cultural survey research (pp. 35–56). Hoboken, NJ: Wiley & Sons.

Harkness JA, Pennell B-A, & Schoua-Glusberg A (2004). Survey questionnaire translation and assessment In Pressler S, Rothgeb JM, Couper MP, Lessler JT, Martin E, Martin J & Singer E (Eds.), Methods for testing and evaluating survey questionnaires (pp. 453–473). Hoboken, NJ: Wiley & Sons, Inc.

Hughes KA (2004). Comparing pretesting methods: Cognitive interviews, respondent debriefing, and behavior coding (No. Research Report Series (Survey Methodology #2004–02)). Washington D.C.: Statistical Research Division, US Bureau of the Census.

IARC. (2009). IARC Handbooks of Cancer Prevention: Tobacco Control. Volume 12. Methods for Evaluating Tobacco Control Policies. Lyon, France: International Agency for Research on Cancer.

Johnson TP (1998). Approaches to equivalence in cross-cultural and cross-national survey research In Harkness JA (Ed.), Cross-cultural Survey Equivalence. Mannheim, Germany: ZUMA.

Johnson TP, Cho Y, Holbrook A, O' Rourke D, Warnecke R, & Chávez N (2006). Cultural variability in the comprehension of health survey questions. Annals of Epidemiology, 16(9), 661–668. [PubMed: 16473526]

Kissam E, Herrera E, & Nakamoto JM (1993). Hispanic response to Census enumeration forms and procedures. Washington, DC: US Census Bureau.

Knowles E, & Condon C (1999). Why people say "yes": A dual-process theory of acquiescence. Journal of Personality and Social Psychology, 77, 379–386.

Miles MB, & Huberman AM (1994). Qualitative data analysis (Vol. 2). London: Sage.

Miller K, Fitzgerald R, Caspar R, Dimov M, Gray M, Nunes C, et al. (2009). Design and analysis of cognitive interviews for cross-national testing. Paper presented at the International Conference on Survey Methods in Multicultural, Multinational, and Multiregional Contexts, Berlin.

Miller K, Mont D, Maitland A, Altman B, & Madans J (in press). Implementation and results of a cross-national, structured-interview cognitive test. Quality and quantity: International journal of methodology.

Pan Y, Craig B, & Scollon S (2005). Results from Chinese cognitive Interviews on the census 2000 Long Form: Language, literacy, and cultural Issues (No. Survey Methodology #2005–09). Washington, DC: US Census Bureau.

Pan Y, & de la Puente M (2005). Census Bureau guideline for the translation of data collection instruments and supporting materials: Documentation of how the guideline was developed. Washington, DC: United States Bureau of the Census, Statistical Division.

Smith P (2004). Acquiescent response bias as an aspect of cultural communication style. Journal of Cross-Cultural Psychology, 35, 50–61.

Smith TW (2004). Developing and evaluating cross-national survey instruments In Pressler S, Rothgeb JM, Couper MP, Lessler JT, Martin E, Martin J & Singer E (Eds.), Methods for Testing and Evaluating Survey Questionnaires (pp. 431–452). Hoboken, NJ: Wiley & Sons, Inc.

Thompson ME, Fong GT, Hammond D, Boudreau C, Dreizen PR, Hyland A, et al. (2006). The methodology of the International Tobacco Control Policy Evaluation Project. Tobacco Control, 15(Supp 3), iii12–iii18. [PubMed: 16754941]

Thrasher JF, Boado M, Sebrié EM, & Bianco E (2009). Smoke-free policies and the social acceptability of smoking in Uruguay and Mexico: Findings from the International Tobacco Control Policy Evaluation (ITC) Project. Nicotine & Tobacco Research, 11, 591–599. [PubMed: 19380383]

Thrasher JF, Chaloupka F, Hammond D, Fong GT, Borland R, Hastings G, et al. (2006). Evaluación de las políticas contra el tabaquismo en países latinoamericanos en la era del Convenio Marco para el Control del Tabaco [Evaluation of tobacco control policies in Latin American countries during the era of the Framework Convention on Tobacco Control]. Salud Pública de México, 48(Supp 1), S155–S166. [PubMed: 17684678]

Thrasher JF, & Johnson TP (2008). Developing and assessing comparable questions in cross-cultural survey research on tobacco In IARC (Ed.), IARC Handbooks of Cancer Prevention: Tobacco Control. Volume 12. Methods for evaluating tobacco control policies. Lyon, France: International Agency for Research on Cancer.

Van de Vijver FJR (2004). Bias and equivalence: Cross-cultural perspectives In Harkness JA, Van de Vijver FJR & Mohler PP (Eds.), Cross-cultural Survey Methods. Hoboken, NH: Wiley & Sons, Inc.

Van de Vijver FJR, & Leung K (1997). Methods and data analysis for cross-cultural research. London: Sage.

Van der Zouwen J, & Smit JH (2004). Evaluating survey questions by analyzing patterns of behavior codes and question-answer sequences: A diagnostic approach In Pressler S, Rothgeb JM, Couper MP, Lessler JT, Martin E, Martin J & Singer E (Eds.), Methods for Testing and Evaluating Survey Questionnaires (pp. 109–130). Hoboken, NJ: Wiley & Sons, Inc.

Van Herk H, Poortinga YH, & Verhallen TMM (2005). Equivalence of survey data: relevance for international marketing. European Journal of Marketing, 39(3/4), 351–364.

WHO. (2003). Framework Convention on Tobacco Control. Geneva, Switzerland: World Health Organization, Tobacco Free Initiative.

Willis GB (2005). Cognitive interviewing. London: Sage.

Willis GB, Lawrence D, Hartman A, Stapleton Kudela M, Levin K, & Forsyth B (2008). Translation of a tobacco survey into Spanish and Asian languages. Nicotine & Tobacco Research, 10(6), 1075–1084. [PubMed: 18584471]

Willis GB, & Zahnd E (2007). Questionnaire design from a cross-cultural perspective: An empirical investigation of Koreans and Non-Koreans. Journal of Health Care for the Poor and Underserved, 18, 190–210.

Zahnd E, Tam T, Lordi N, Willis GB, Edwards S, Fry S, et al. (2005). Cross-cultural behavior coding: Using the 2003 California Health Interview Survey (CHIS) to assess cultural /language data quality. Paper presented at the European Association for Survey Research, Barcelona, Spain.

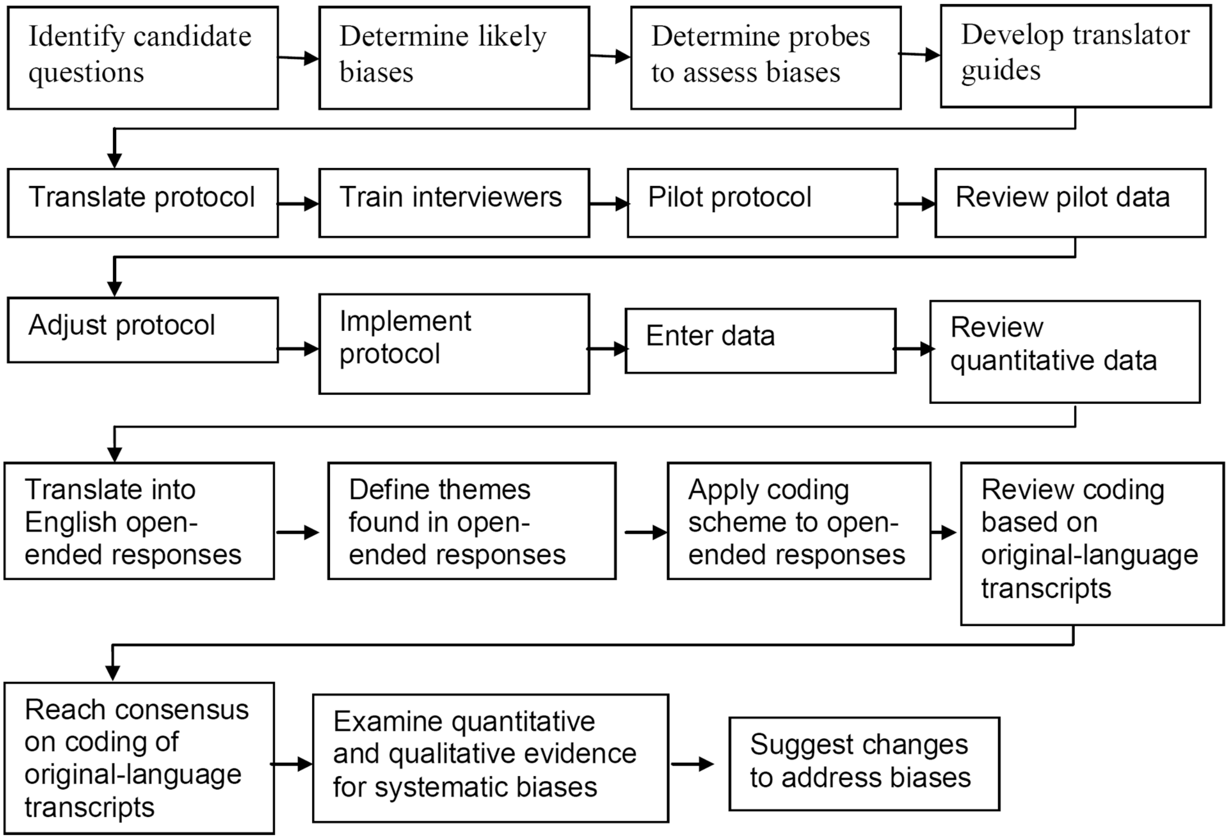
Identify candidate questions
Determine likely biases
Determine probes to assess biases
Develop translator guides
Translate protocol
Train interviewers
Pilot protocol
Review pilot data
Adjust protocol
Implement protocol
Enter data
Review quantitative data
Translate into English open-ended responses
Define themes found in open-ended responses
Apply coding scheme to open-ended responses
Review coding based on original-language transcripts
Reach consensus on coding of original-language transcripts
Examine quantitative and qualitative evidence for systematic biases
Suggest changes to address biases

Figure 1.
Protocol development steps

Table 1.

Sample characteristics

		Australia (n=20)	US (n=20)	Uruguay (n=20)	Mexico (n=20)	Malaysia (n=20)	Thailand (n=20)
Male		50%	65%	50%	65%	100%	80%
Age		36	36	40	38	31	39
Education	*<HS*	30%	5%	30%	20%	10%	85%
	HS	30%	65%	40%	55%	60%	10%
	Uni	40%	30%	30%	25%	30%	5%
Daily smoker		90%	85%	80%	85%	90%	85%
Average cigs/day		14.0	19.6	12.9	14.2	12.9	12.7

Table 2.

Behavioral coding of responses to three example questions

Question	Behavior	AU N (%)	US N (%)	UY N (%)	MX N (%)	MY N (%)	TH N (%)	Overall N (%)	Fisher's Test
Of the five closest friends or acquaintances that you spend time with on a regular basis, how many of them are smokers?	Needed question repeated	3 (15%)	2 (10%)	0 (0%)	1 (5%)	4 (20%)	3 (15%)	13 (11%)	0.349
	Had difficulty using the response options	1 (5%)	8 (40%)	0 (0%)	1 (5%)	2 (10%)	6 (30%)	18 (15%)	0.001
	Asked for clarification/qualified answer	2 (10%)	2 (10%)	0 (0%)	0 (0%)	4 (20%)	1 (5%)	9 (8%)	0.155
Is the following statement true or false? The way a smoker puffs on a cigarette can affect the amount of tar and nicotine a smoker takes in.	Needed question repeated	1 (5%)	1 (5%)	0 (0%)	0 (0%)	1 (5%)	1 (5%)	4 (3%)	1.000
	Had difficulty using the response options	3 (15%)	6 (30%)	0 (0%)	0 (0%)	1 (5%)	1 (5%)	11 (9%)	0.007
	Asked for clarification/qualified answer	2 (10%)	6 (30%)	0 (0%)	1 (5%)	1 (5%)	0 (0%)	10 (8%)	0.009
Tobacco is addictive (5-point Likert)	Needed question repeated	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	1 (5%)	1 (1%)	1.000
	Had difficulty using the response options	1 (5%)	1 (5%)	0 (0%)	0 (0%)	1 (5%)	1 (5%)	4 (3%)	1.000
	Asked for clarification /qualified answer	0 (0%)	1 (5%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	1 (1%)	1.000

Note: n=20 in each country

Table 3.

Frequency of themes in descriptions of addiction

Themes	AU N (%)	US N (%)	UY N (%)	MX N (%)	MY N (%)	TH N (%)	Overall N (%)	Fisher's Test
Control-General	10 (50%)	10 (50%)	11 (55%)	11 (55%)	9 (45%)	6 (30%)	57 (48%)	0.602
Control-Physical	5 (25%)	8 (40%)	1 (5%)	3 (15%)	0 (0%)	0 (0%)	17 (14%)	0.000
Control-Psych	2 (10%)	2 (10%)	0 (0%)	0 (0%)	6 (30%)	0 (0%)	10 (8%)	0.002
Control-Quit	2 (10%)	0 (0%)	1 (5%)	2 (10%)	2 (10%)	3 (15%)	10 (8%)	0.723
Frequency-Quantity	10 (50%)	1 (5%)	8 (40%)	3 (15%)	4 (20%)	6 (30%)	32 (27%)	0.013
Danger	0 (0%)	1 (5%)	0 (0%)	0 (0%)	0 (0%)	7 (35%)	8 (7%)	0.000
Pleasure	0 (0%)	1 (5%)	1 (5%)	0 (%)	1 (5%)	1 (5%)	4 (3%)	1.000

Note: n=20 in each country

Table 4.

What are some other things that are addictive?

Responses	AU N (%)	US N (%)	UY N (%)	MX N (%)	MY N (%)	TH N (%)	Overall N (%)	Fisher's Test
Alcohol	15 (75%)	16 (84%)	12 (63%)	11 (73%)	1 (7%)	4 (22%)	59 (56%)	0.000
Drugs	15 (75%)	15 (79%)	14 (74%)	11 (73%)	2 (14%)	15 (83%)	72 (68%)	0.000
Pleasurable foods	4 (20%)	8 (42%)	6 (32%)	3 (20%)	2 (14%)	0 (0%)	23 (22%)	0.015
Sex	2 (10%)	3 (16%)	0 (0%)	4 (27%)	1 (7%)	0 (0%)	10 (10%)	0.120
Gambling	3 (15%)	1 (5%)	2 (11%)	1 (7%)	0 (0%)	0 (0%)	7 (7%)	0.417
Other	3 (15%)	9 (47%)	4 (21%)	2 (13%)	8 (57%)	1 (6%)	27 (26%)	0.010

Note: n=20 in each country