



Published in final edited form as:

Nicotine Tob Res. 2008 June ; 10(6): 1075–1084. doi:10.1080/14622200802087572.

Translation of a Tobacco Survey into Spanish and Asian Languages: The Tobacco Use Supplement to the Current Population Survey

Gordon Willis, PhD,

Division of Cancer Control and Population Sciences, National Cancer Institute, National Institutes of Health

Deirdre Lawrence, PhD,

Division of Cancer Control and Population Sciences, National Cancer Institute, National Institutes of Health

Anne Hartman, MS,

Division of Cancer Control and Population Sciences, National Cancer Institute, National Institutes of Health

Martha Stapleton Kudela, MA,

Westat

Kerry Levin, PhD, and

Westat

Barbara Forsyth, PhD

Westat

Abstract

Because of the vital need to attain cross-cultural comparability of estimates of tobacco use across subgroups of the U.S. population that differ in primary language use, the NCI Tobacco Use Special Cessation Supplement to the Current Population Survey (TUSCS-CPS) was translated into Spanish, Chinese (Mandarin and Cantonese), Korean, Vietnamese, and Khmer (Cambodian). The questionnaire translations were extensively tested using an eight-step process that focused on both translation procedures and empirical pretesting. The resulting translations are available on the Internet (at <http://riskfactor.cancer.gov/studies/tus-cps/translation/questionnaires.html>) for tobacco researchers to use in their own surveys, either in full, or as material to be selected as appropriate. This manuscript provides information to guide researchers in accessing and using the translations, and describes the empirical procedures used to develop and pretest them (cognitive interviewing and behavior coding). We also provide recommendations concerning the further development of questionnaire translations.

In order to estimate patterns of tobacco use in the United States, it is vital to assess behavior accurately across a range of racial and ethnic minority groups. Especially because many members of these sub-populations are linguistically isolated, and cannot complete a questionnaire in English, researchers must develop questionnaires that are both linguistically

Corresponding author: Gordon Willis, ARP/DCCPS/NCI/NIH, 6130 Executive Blvd, MSC 7344, EPN 4005, Bethesda, MD, 20892-7344, USA, (301) 594-6652, (301) 435-3710 (fax), willisg@mail.nih.gov.

The research described in this manuscript was performed at the National Cancer Institute, NIH, and at Westat.

No author of this manuscript has a conflict of interest with respect to the research reported.

and culturally appropriate (Behling & Law, 2000; Goerman, 2006; Ji, Schwarz, & Nisbett, 2000; Johnson, 1988, 2006; Harkness & Schoua-Glusberg, 1998; Martinez, Marín, & Schoua-Glusberg, 2006; McKay, et al., 1996; Rogler, 1999; Stewart & Nápoles-Springer, 2000; Warnecke, et al., 1997). Several large population surveys, such as the California Health Interview Survey (CHIS), have begun to include a range of languages (Ponce, et al., 2004). Further, translated versions are especially appropriate for tobacco use ascertainment in State and community-level areas that are highly represented by non-English-speaking populations (Ma, et al., 2004).

The current research focuses on the translation of a well-established tobacco use questionnaire -- the Tobacco Use Supplement to the Current Population Survey (TUS-CPS) -- into Spanish and several Asian languages. We direct users to versions of the TUS-CPS translated questionnaires that are downloadable from the Internet, and advocate their use in National, State, and local area surveys. Further, we describe the extensive processes used to develop the translations, because an explication of the methodology may be of interest to researchers who plan their own translation and evaluation efforts.

Background

As a general population surveillance system, the TUS-CPS has been administered six times since 1992¹. The survey is a key source of National and State level data on smoking and other tobacco use in the U.S. household population, and for each cycle relies on a large, nationally representative sample of about 240,000 individuals age 15 or higher. Data from the TUS-CPS are used by researchers to monitor progress in the control of tobacco use, to conduct tobacco-related research, and to evaluate tobacco control programs (Augustson & Marcus, 2004; Burns, Major, Anderson, & Vaughn, 2003; Burns & Warner, 2003; Levy, Romano, & Mumford, 2005; National Cancer Institute, 2005; Rivara, et al., 2004). Approximately 75% of interviews are conducted by telephone, and 25% through in-person interview.

To obtain information that extends beyond the individual respondent's current tobacco use, and to characterize and monitor tobacco cessation practices, NCI developed and first fielded in 2003 an elaborated version of the base TUS-CPS: The Tobacco Use Special Cessation Supplement to the Current Population Survey (the TUSCS-CPS). The TUSCS-CPS focuses on a range of cessation-related behaviors, and contains items on the topics listed in Table 1. The Cessation version, as administered in 2003, was selected as the instrument to be translated and evaluated in the effort described here. We first indicate how to access the translations, and then review the methods used in their development.

Accessing and using translated TUSCS-CPS questionnaires

The language translations are downloadable directly from <http://riskfactor.cancer.gov/studies/tus-cps/translation/questionnaires.html>. Translated versions are intended mainly for telephone-based administration, but can also be adapted easily to in-person interviewing, and with somewhat more modification, to self-administration. Files are provided in Adobe Pro format, which can be edited as the researcher desires. We caution, however, that modifications to survey question wording or ordering can have influences on responses that are difficult to predict (Tourangeau, Rips, & Rasinski, 2000). Each translated questionnaire is available as both a computer-assisted interview (CAI) instrument and a paper-and-pencil

¹The TUS-CPS is primarily sponsored by the National Cancer Institute (NCI); the Centers for Disease Control and Prevention (CDC) has been a co-sponsor since 2001-02; the CPS is itself a joint effort of the Bureau of Labor Statistics and the Census Bureau (see <http://www.census.gov/cps/>)

interview version (PAPI). The CAI instrument contains detailed specifications for question programming and sequencing (skip patterns), and interviewer instructions.

For the PAPI instruments, all programming language has been deleted or simplified such that skip patterns and administration instructions can be read easily by the interviewer. Also, some automated range checks that appear in the CAI instrument have been removed from the paper version. English and Spanish versions of the online TUSCS-CPS questionnaires are slightly different from those administered in 2003, in part to allow the current version to function as a stand-alone instrument not dependent on the Core CPS, and partly to incorporate limited modifications to several questions, based on the results of testing the translations².

Development of the TUSCS-CPS Translations

Rationale for translation into Asian languages

Although the 2003 TUSCS-CPS was fielded in English and Spanish, NCI made a further decision to create an instrument for more widespread use, by converting the questionnaire into a set of translations for several Asian groups: Korean, Chinese (Mandarin and Cantonese), Vietnamese, and Cambodian (Khmer)³. Translation into Asian languages is particularly important, given that measurement of tobacco use among U.S. Asian sub-populations has proven enigmatic (Kim, Ziedonis, & Chen, 2007). In particular, although large National surveys have sometimes revealed lower rates of smoking for Asian Americans (American Lung Association, 2005; Centers for Disease Control and Prevention, 1998), estimates for these groups may be error prone (Kim, et al., 2007). Prevalence estimates are often based on surveys of Asians who speak English, and who may exhibit smoking characteristics very discrepant from those of recent immigrants who are less acculturated, and speak only their native language. For example, Ma, et al. (2004) and Kim, et al. (2007) found that low-aculturated Asian men reported smoking rates that were substantially higher than those of high-aculturated Asian men who spoke English.

Challenges to translation and establishment of cross-cultural comparability

Although the utility of translated questionnaire instruments is clear, it is generally insufficient to simply translate the source version through a straightforward, single-step translation process. Rather, survey translation is a complex endeavor requiring the development and implementation of a number of careful practices (Brislin, 1970; Census Bureau, 2004; Harkness & Schoua-Glusberg, 1998; Harkness, Van de Vijver & Mohler, 2003; McKay, et al., 1996; Ponce, et al., 2004). Most vexing are the severe challenges to creating survey questionnaires that exhibit the key characteristic of *cross-cultural comparability* of results (Miller, 2004; Nápoles-Springer & Stewart, 2006; Pan & de la Puente, 2005; Schmidt & Bullinger, 2003; Singelis, et al., 2006; Yu, Lee, & Woo, 2004). Although researchers have generated a multitude of conceptualizations of the construct of comparability (or equivalence), a common theme is that the ideas conveyed in one language may be easily distorted through translation procedures that do not take into account both linguistic and cultural variation between groups. As an example, Nápoles-Springer, Santoyo-Olsson, O'Brien, & Stewart (2006) reported that the phrase “medical tests and procedures” failed to function well for Latinos, as it did not clearly bring to mind particular events. The authors concluded that several examples (blood test, x-ray, cancer screening tests) should be added to all questionnaire versions to enhance comprehension and resultant comparability.

²See <http://riskfactor.cancer.gov/studies/tus-cps/translation/researchers.html> for details.

³The TUSCS-CPS translation into Cambodian (Khmer) has not undergone extensive review or empirical testing as the other versions, but is offered for use as well.

For several decades, survey researchers have sought cross-cultural comparability by following procedures described by Brislin (1970), involving forward translation from the source to the target language, and then independent back-translation to the original (e.g., Yu, et al., 2004). The original and back-translated versions are compared, and discrepancies noted and reconciled. However, recent reviews by survey researchers and linguists have concluded (as did Brislin's original article) that back-translation should not be regarded as a singular "best practice" (Ponce, et al, 2004)⁴. In particular, a tendency towards over-reliance on word-for-word, literal translations may sacrifice the overall intent of the translated item (Carlson, 2000; Census Bureau, 2004; European Social Survey, 2002; Forsyth, Kudela, Levin, Lawrence, & Willis, 2007; Harkness, et al., 2003; Harkness & Schoua-Glusberg, 1998; Ponce, et al, 2004)⁵.

As an alternative, researchers have increasingly advocated a multi-step approach to translation (Census Bureau, 2004). Table 2 summarizes the approach taken in the current investigation. Instead of back translation, this approach relies initially on an expert-team approach consisting of translation and review by a research team whose members are not only bilingual but also knowledgeable about questionnaire design and the measurement objectives of each survey question. Rather than producing literal translations that may retain only individual-word-level comparability, a translate-and-review team strives to maintain overall question intent.

Beyond the translation process, Harkness, et al. (2003) and the Census Bureau (2004) also recommend steps devoted to quality assurance of the translated versions, involving empirical evaluation and pretesting. Empirical testing of self-report instruments has a venerable tradition within the psychological and sociological fields, and often incorporates concepts from psychometrics such as scale reliability (e.g., Cronbach's alpha) and Item Response Theory (DeVellis, 2003; Reeve, 2005). Such psychometric approaches have been used successfully to evaluate the cross-cultural comparability of self-report instruments (Singelis, et al., 2006). However, a major limitation of psychometrics is their focus on latent constructs as measured through the use of scales which rely on several items in conjunction to tap each construct. The psychometric approach is not designed to be applicable to observable, "platonic measures" involving single-item approaches to behaviors like tobacco use (DeVellis, 2000; Willis, 2005).

Accordingly, the TUS-CPS questions were evaluated through the adaptation of survey pretesting techniques -- cognitive interviewing and behavior coding -- that are commonly used to evaluate monolingual survey questions (Forsyth & Lessler, 1991; Willis, 2005). Cognitive interviewing involves intensive interviews of small, purposive samples of respondents ("subjects"), who are administered the evaluated questionnaire, and also verbally probed by a specially trained interviewer to elucidate problems in comprehension, recall, decision-making processes, or response category selection, that are otherwise not evident. For example, the question "*Would you say your health in general is excellent, very good, good, fair, or poor?*" is followed by the probe "*Why do you say that your health is (response)?*". Increasingly, cognitive interviews are being used to evaluate issues of cultural comparability of survey instruments (Agans, Deeb-Sossa, & Kalsbeek, 2006; Forsyth, et al., 2007; Martinez, et al., 2006; Miller, 2004; Nápoles-Springer & Stewart, 2006; Nápoles-

⁴See Yu, et al. (2004) for a dissenting view. One complication is that back-translation as a single step has increasingly been criticized -- yet neither Brislin nor many later advocates of back-translation advocated use of this technique alone, and instead favored a multi-step translation and evaluation process.

⁵Rogler (1999) provides a compelling example to make this point: The phrase "feeling blue" might be translated from English into Spanish, and then "successfully" back-translated to produce the original English wording. However, as the Spanish word for blue, *Azul*, carries no meaning as a descriptor of mood, the Spanish version is nonsensical.

Springer, et al., 2006; Pasick, Stewart, Bird, & D'Onofrio, 2001; Quittner, et al., 2000; Willis, 2005).

A second well-established pretesting technique is behavior coding, which focuses on the overt behaviors of the interviewer and the respondent as they interact during the (interviewer-administered) survey interview (Cannell, Fowler, & Marquis, 1968; Fowler, 1995; Fowler & Cannell, 1996). Unlike cognitive interviewing, behavior coding is a passive endeavor, in which a specially trained coder evaluates recordings of interviews to assess interviewer-respondent interactions, and introduces no probing into the ongoing interview. Interactions are systematically coded with respect to behaviors that may be indicators of difficulties that adversely affect data quality (e.g., the respondent demonstrates difficulty in understanding the question). Because it is a quantitative endeavor, behavior coding provides aggregate code summaries to represent the magnitude of a particular problem. On the other hand, behavior coding provides less opportunity than cognitive interviewing to locate covert "silent misunderstandings" (DeMaio & Rothgeb, 1996). Like cognitive interviewing, behavior coding has increasingly been applied within the cross-cultural domain (Edwards, Levin, Willis, Lawrence, & Thompson, 2005; Hunter & Landreth, 2006; Zahnd, et al., 2005).

For current purposes, we focus mainly on the results of pretesting obtained through cognitive interviewing and behavior coding steps, as opposed to the details of the TUSCS-CPS translation procedures (a full description is provided by Forsyth, et al., 2007). The results of these steps are informative for two reasons: (1) They illustrate how the adaptation of these techniques can be used in the cross-cultural domain, and (2) They reveal the types of problems that would have been missed, absent these techniques.

Cognitive Interviewing Step

Method⁶

A total of 41 cognitive interviews (Table 2, Step 4) followed initial translation activities, and involved adult (18 or older) speakers of Spanish (9 interviews), Chinese (9 interviews each of Mandarin and Cantonese), Korean (9 interviews), and Vietnamese (14 interviews), who were recruited from the local community on the basis of racial/ethnic group membership and primary language use⁷. One-hour, face-to-face interviews were conducted in a cognitive research laboratory by contractor (Westat) staff or by subcontractors selected based on specific language characteristics. Interviews incorporated verbal probing methods described by Willis (2005), involving one-on-one interviews conducted in a private location. Interviewers used a protocol containing probe questions designed to identify problems with the survey questions⁸ (e.g., "*In your own words, what is 'snuff'?*").

Results

Cognitive interviews revealed few problems in Spanish-language interviews; issues that arose largely centered on subtleties of translation. For example, a set of response categories used across several questions - 'Nunca', 'Un poco', 'Algo', and 'Muy' (roughly 'not at all,' 'a little', 'somewhat', and 'very') - were modified so that the term 'cierta' (sure) was added to each term for clarity. Asian-language cognitive interview results presented a somewhat different picture. Despite our careful attention to translation (in Steps 1–3, Table 2), Asian

⁶See <http://riskfactor.cancer.gov/studies/tus-cps/translation/reports.html> for complete information.

⁷Cognitive interviews of the TUSCS-CPS have previously also been conducted in English, by Census Bureau staff. Those results were used as a reference point for assessing the current non-English versions.

⁸Note that probe questions themselves were translated into non-English languages, requiring careful attention to translation of not only the targeted survey items, but as well of the probes used in the evaluation of those items.

translations revealed a range of problems, sometimes involving potentially serious sources of misunderstanding. For example, testing of the Korean version revealed that the translation of a question asking if anyone smoked at home had reversed the phrasing, relative to English, and was expressed as: “*Is it true that there is no one who smokes cigarettes, cigars, or pipes anywhere inside of your home?*” As such, the correct answer if no one smokes is “yes” (i.e., “*Yes, it’s true that no one smokes cigarettes*”). Because this not only proved confusing, but could have resulted in responses that were logically reversed between Korean and English interviews, the Korean question wording was revised to match the English.

As an example of a more subtle translation-oriented problem, cognitive interviews in Vietnamese revealed that within the context of the question “*In your opinion, how easy is it for minors to buy cigarettes and other tobacco products in your community?*” the word “community” tended to be interpreted as the Vietnamese people in general. As an alternative, the Vietnamese word for “neighborhood” was considered but also felt by Vietnamese-speaking staff to be problematic, as it implies that the respondent has a relationship with neighbors. We therefore chose the Vietnamese equivalent of “*In your opinion, how easy is it for a minor to buy cigarettes or cigarette products in the area where you are now living?*” Finally, as an example of an inherent knowledge problem revealed through cognitive testing, we found that a question asking whether respondents have ever switched from a stronger to a lighter cigarette was found to pose difficulties for respondents who had smoked Korean or Chinese brands that are unlabelled with respect to tar or nicotine content.

Following the cognitive interviews, the translations were modified to address such identified problems when deemed possible (See Table 2, Step 5: Second Adjudication). In general, problems related to translation errors could be remedied, whereas those due to inherent limitations in knowledge (such as that involving cigarette strength) were not possible to address through question rephrasing, and were deferred to the following, Behavior Coding step.

Behavior Coding Step

Method

Sample—Following cognitive interviewing, the research team conducted a pilot test (Table 2, Step 6) that was of significant size to collect quantitative information appropriate for behavior coding analysis. A total of 418 interviews were completed of individuals who had smoked at least 100 cigarettes in their lifetime: 125 in English, 51 in Spanish, 57 in Mandarin, 67 in Cantonese, 66 in Korean, and 52 Vietnamese, across 39 U.S. States. No efforts were made to control demographic characteristics, or to produce statistically weighted data, because for purposes of questionnaire development and evaluation a key objective is to obtain as wide a variety of participant characteristics as possible (Willis, 2005). Participant selection was achieved through several means, including (a) random digit dialing (RDD); (b) a list sample consisting of households containing a member speaking one of the five (non-English) languages; (c) a list of surnames that are usually representative of Hispanic and Asian groups; and (d) (twelve) participants identified by screened household members who suggested a potentially-eligible acquaintance.

Table 3 illustrates selected demographic characteristics from the pilot test.⁹ Mean age of participants (48.6 years) varied significantly between groups ($p < .05$, by ANOVA), from

⁹Most demographic distributions were unsurprising. For example, virtually all participants interviewed in Asian languages reported having an Asian, non-Hispanic background, and all interviewed in Spanish reported Hispanic ethnicity. However, 30.4% of those completing the questionnaire in Spanish indicated an Asian racial background. We are unable to explain this unanticipated result.

41.4 for English, to 50.9 for Spanish, with Asian groups intermediate; however, group variances were roughly similar (standard deviations ranged from 11.7 to 15.2). Gender distribution was very discrepant across groups ($p < .05$, by Chi-square test): English and Spanish were adequately represented by female participants (34.4% and 39.2%, respectively), but no Asian group contained more than 10.6% females. Our failure to successfully recruit female Asian smokers is consistent with the observation that self-reported smoking by Asian females is low, compared to men (Kim, et al., 2007). With respect to the key variable of current smoking status, the groups were fairly equally divided, with 75.6% current everyday smokers, 12.7% some-day smokers, and 11.7% former smokers. The mean number of cigarettes smoked per day for current every-day smokers was 15.0, and ranged from 12.2 for Vietnamese, to 18.2 for English-speakers (differences in means varied significantly across groups ($p < .001$, by ANOVA).

Interestingly, pilot test interviews took substantially longer ($p < .05$, by ANOVA) in Vietnamese (mean of 39.4 minutes) than in any other language (for which means ranged from 26 to 31 minutes). A language consultant suggested that Vietnamese have a tendency to engage in a significant degree of verbalization, even for yes-no survey answers, but this is speculative. Further, interview length varied from 10 to 100 minutes. In general, researchers should anticipate that surveys in languages other than English will likely take longer than the English version.

Procedures for behavior coding—Behavior codes used within the current investigation are contained in Table 4.

From the TUSCS-CPS questionnaire, 35 of the items that were considered most important (because they either provide key prevalence estimates, or control branching to further major item sets) were analyzed via behavior coding; 27 of these items produced sample sizes appropriate for quantitative analysis (we chose, as a cutoff value, 20 or more observations per cell). In brief, specially trained bilingual coders (who were not the original interviewers) reviewed digital recordings of interviews, and assigned codes as appropriate to each interviewer-participant interchange. Subsequent to coding, coders were debriefed by project staff, to lend a qualitative element to their assessment of item functioning.

Results

Concerning the manner in which both interviewers and participants were observed to react to each tested question, we present detailed quantitative results of the behavior coding at <http://riskfactor.cancer.gov/studies/tus-cps/translation/reports.html>. Here, we provide key examples of issues that emerged from this activity. Each of these was considered a potential threat to data quality, and prompted a consideration of modifications to the instrument, and will be described in turn, according to a four-category system developed for this investigation.¹⁰

Interviewer-based problems were those that emerged mainly as a function of the interviewer's Reading Errors, which in turn may induce difficulties for participants. For the 27 evaluated questions, the mean percentage of interactions that produced Reading Errors was 6.4%, in English, 15.3% in Spanish, 16.5% in Mandarin, 16.0% in Cantonese, 12.5% in Korean, and 16.4% in Vietnamese. Despite this variation, we concluded that questions were generally readable as written, with several notable exceptions that prompted question revision. Correlation analysis that relied on the mean frequency of Reading Errors for each

¹⁰We recognized that it is sometimes difficult to unambiguously assign a coding category to each observed problem. Our major intention in using this scheme is simply to provide a heuristic device for organizing our detailed examples into a set of general categories of results.

question illustrated a high correlation between English and Spanish versions ($r = .71, p < .001$), and between English and Vietnamese ($r = .60, p < .001$), indicating that questions causing problems for interviewers in English also did so in Spanish and in Vietnamese.

However, the corresponding correlations for all other language-pair comparisons were non-significant, and interestingly, the correlation between Cantonese and Mandarin Reading Errors was only .30, despite the use of a shared character system between these languages. This finding suggests that a great deal of the variation between versions is likely due to divergent interviewer compliance in reading the questions, as opposed to inherent characteristics of the languages per se. For example, some interviewers failed to read the instruction “*Please tell me if each of the following is true for you*” before asking “*You have trouble going for more than a few hours without smoking,*” which created difficulties for participants who were left unsure of how to respond.

To assess whether such Reading Errors did produce problems for participants, we correlated the frequency of Reading Errors with the frequency with which the participant in turn produced any behavior code indicating a problem, over all coded items. For English language interviews, we observed a high correlation between Reading Errors and participant behavior codes ($r = .61, p < .01$), suggesting that when the interviewer misread the item, this did produce problems for participants. For each of the other languages, the relationship between Reading Errors and participant-oriented codes was non-significant, and sometimes near zero. This finding suggests that interviewer misreading may have not always been detrimental, but rather produced a mixture of help and harm to participants, to varying degrees (perhaps as the interviewer attempted to unilaterally improve the wording). Overall, these analyses suggest important issues related to interviewer training that will be discussed later.

A second major problem category, *translation problems*, involved items that failed to express the meaning as intended to participants due to defects in the translation process. Translation problems were diagnosed through discussion with bilingual interviewers and behavior coders. For example, the question: “*What price did you pay for the LAST pack/carton of cigarettes you bought? - Please report the cost after using discounts or coupons*” was mistranslated into Chinese such that it directed the participant to report *only* the cost of the coupon (the opposite of the intended meaning). The problem was reflected by an overall high frequency of codes associated with problems answering the question for Chinese participants (34% in Cantonese and 32% in Mandarin, as opposed to 14% in English). As a response, the translation was modified to express the intended concept.

As a particularly interesting translation problem, the seemingly clear question “*Have you smoked at least 100 cigarettes in your entire life?*” produced requests for clarification for 35.8% of interviews in Cantonese and 34.8% in Vietnamese, but only 4.3% in English. Debriefing of interviewers and coders determined that the translation of the term “*in your entire life*” meant literally “*from birth to death*” in several Asian languages, and therefore required participants to speculate about future behavior, rather than only reporting past experiences.¹¹ The phrase was re-translated to better match the English version, and to better convey the notion “*until now.*”

A third set of difficulties, *problems of cultural adaptation*, supersede translation issues, as they were not clearly related to literal mis-translation. As examples:

¹¹This item also created problems of item sensitivity, as we were informed that for some Asian cultures it is considered unlucky to contemplate specifically the circumstances of one’s own death, a consideration that was triggered by our original translation.

1. Very few Asian participants were familiar with snuff (however translated). A definition was therefore added for use at the interviewer's discretion;
2. A four-item scale on nicotine dependence presented a statement and then asked whether it is true or not (e.g., "*When you go without smoking for a few hours, you experience craving.*") This format tended to produce long explanations, as opposed to the desired true/false response, from Mandarin Chinese in particular. A Chinese cognitive interviewer suggested that this occurred due to unfamiliarity with the cultural conventions associated with answering questionnaire items (i.e., that the desired response is a simple "yes" or "no", as opposed to an explanation that is more appropriate for naturalistic conversation). To clarify and direct the response task, we added the instruction: "*Please tell me if each of the following statements is true for you. You may answer with true or false, or with yes or no*";
3. Asian participants tended to produce high frequencies of behavior codes indicating problems with providing a response on a 0 to 10 scale. Interviewers and behavior coders suggested that Asians are not familiar with the use of this scale. Because our intent was to maintain similarity to the English version, and because the pilot test demonstrated that Asians had little difficulty with the scale once the interviewer further explained it, we added a standardized instruction to Asian versions: "*Overall, on a scale from 1 to 10, where 1 is not at all interested and 10 is extremely interested, how interested are you in quitting smoking. Please indicate how interested you are in quitting by picking a number from 1 to 10.*"

As a final problem category, we observed *generic design problems* that were not specific to any particular tested group. For example, the question "*How soon after you wake up do you typically smoke your first cigarette of the day?*" tended to produce the same reaction for every version (including English): Rather than answering with a quantitative value, such as "one hour," roughly half the participants reported in terms like "before breakfast" or "as soon as I open my eyes." In some cases of this type, bilingual translators could offer no suggestions for modification, and the translated version was unmodified. Within a field survey environment, administrators should therefore expect to receive imprecise responses for this item that require further interviewer probing, for any respondent group.

Discussion

Based on our eight-step developmental approach, we were able to identify and in many cases remediate problems that were rooted in interviewer (mis)behavior, translation error, problems of cultural adaptation, and general problems of question design. Based on the nature of these problems, we can suggest several conclusions concerning the nature of translated instruments. First, for interviewer-administered instruments, we suggest that data quality is not simply a function of correct translation, but also of proper delivery. Even where questions were found to contain no obvious translation problems, interviewers sometimes misread questions in ways that fundamentally affected their meaning. In training, researchers need to stress that interviewers should read each question as worded, to avoid variance in responses due only to discrepancies in interviewer performance. Or, if interviewers are allowed to deviate from standard wording (in an attempt to assist respondents), it is important to carefully monitor interviews for signs of bias.

Second, our results suggest that respondents answering the questionnaire in languages other than English, especially if elderly or unacculturated to American society, may not be as familiar with the implicit rules of questionnaire interaction. They may therefore be less likely to choose the specific response categories provided, or to answer in the terms implied by the question. The interviewer in these cases needs to be prepared to use neutral follow-up

probing to elicit codeable responses (e.g., “How many minutes or hours would you say it is before you smoke your first cigarette of the day?”).

Further, we note that even an intensive translation effort relying on multiple experts was not entirely successful in eradicating some clear errors subsequently identified through empirical pretesting, and that in general, each evaluation step served as a successive filter to catch problems that had slipped through earlier steps. However, despite the use of a multi-step process, some problems may be persistent, such that sources of cross-cultural non-comparability sometimes cannot be eradicated through the discovery of a corrected verbal translation. In particular, a basic concept (e.g., “light” versus “regular” cigarette) may not be represented in such a way that survey questions concerning that concept will be meaningful for all groups. However, by pretesting we can at least identify these limitations, and be prepared to interpret results with important caveats.

Alternately, the pretesting process may reveal particular problems that can be resolved – but that may also require modification of the source (e.g., English) questionnaire, especially when the problems appear generic in nature and not tied to a particular cultural group or translated version. Therefore, for investigations in which constraints on the source version are less severe than for the TUSCS-CPS, we advocate the practice of *decentering* (McKay, et al., 1996; Werner & Campbell, 1970), in which the original questionnaire is considered to be open to modification, based on results obtained through testing of the translations.

Caveats and suggestions for further research

First, we acknowledge that the techniques used – cognitive interviewing and behavior coding – are relatively new in application to cross-cultural research, and can potentially be applied in a wide variety of ways. As such, there are no commonly accepted practices known to guarantee validity and to avoid biased results. Most seriously, given that assessment of difficulties across languages necessarily involved different sets of bilingual researchers, our cognitive probing or coding methods might have been applied inconsistently across group, such that observed discrepancies between language versions therefore reflect between-group variation in our testing processes. On the other hand, we did find that our interviewers independently recorded difficulties that were common across languages (generic problems), which is somewhat reassuring, as it represents agreement concerning the identification of a common subset of problems with the evaluated items.

A second limitation is that our pilot test, which served as the basis for behavior coding, failed to obtain equivalent participant groups, with respect to demographic variables that might influence the survey response process. In particular, the significance of the wide gender discrepancies across our groups is unclear – future studies might consider more intensive efforts to recruit female Asian smokers in particular. As a final caveat, our measures of item function were indirect, as we lacked criterion measures by which to directly assess degree of response error produced by the evaluated survey questions, for any language group. We therefore advocate future research efforts that access true criterion measures (e.g., objective counts of cigarettes smoked in the past 30 days).

Despite these limitations, we conclude that pretesting with cognitive interviewing and behavior coding likely reveals significant defects in translated versions, and appears to be potentially useful as a general set of procedures for maintaining quality control. This investigation therefore supplements the burgeoning literature indicating that it is worth additional time and expense to apply pretesting techniques in pursuit of cross-cultural comparability of translated survey items. In closing, we point out that the field of survey translation and cross-cultural evaluation is dynamic, and awaits the establishment of demonstrated “best practices.” We suggest that researchers attend to further developments in

the relevant research area, as these will guide future development and quality assurance of health survey questionnaires generally.

Acknowledgments

The authors would like to thank four anonymous reviewers who provided valuable comments on previous versions of the manuscript. This research was supported through contract N02-PC-54407 between the National Cancer Institute (NIH) and Westat.

References

- Agans RP, Deeb-Sossa N, Kalsbeek WD. Mexican immigrants and the use of cognitive techniques in questionnaire development. *Hispanic Journal of Behavioral Sciences*. 2006; 28:209–230.
- American Lung Association. Lung disease data in culturally diverse communities: 2005. American Lung Association; 2005. Retrieved March 5, 2007 from <http://www.lungusa.org/site/pp.aspx?c=dvLUK900E&b=308853&printmode=1>
- Augustson E, Marcus S. Use of the current population survey to characterize subpopulations of continued smokers: a national perspective on the "hardcore" smoker phenomenon. *Nicotine and Tobacco Research*. 2004; 6:621–629. [PubMed: 15370158]
- Behling, O.; Law, KS. *Translating questionnaires and other research instruments: Problems and solutions*. London: Sage; 2000.
- Brislin RW. Back-translation for cross-cultural research. *Journal of Cross-Cultural Psychology*. 1970; 1:185–216.
- Burns, DM.; Major, JM.; Anderson, CM.; Vaughn, JW. Changes in cross-sectional measures of cessation, numbers of cigarettes smoked per day, and time to first cigarette -- California and national data. In: Ruppert, W.; Amacher, RH.; Marcus, SE.; Shopland, DR., editors. *Those who continue to smoke: Is achieving abstinence harder and do we need to change our interventions?*. Bethesda, MD: U.S. Department of Health and Human Services, Public Health Service, National Institutes of Health, National Cancer Institute; 2003. p. 101-125. (Smoking and Tobacco Control Monograph No. 15., NIH Pub. No. 03-5370)
- Burns, DM.; Warner, KE. Smokers who have not quit: Is cessation more difficult and should we change our strategies. In: Ruppert, W.; Amacher, RH.; Marcus, SE.; Shopland, DR., editors. *Those who continue to smoke: Is achieving abstinence harder and do we need to change our interventions?*. Bethesda, MD: U.S. Department of Health and Human Services, Public Health Service, National Institutes of Health, National Cancer Institute; 2003. p. 11-31. (Smoking and Tobacco Control Monograph No. 15., NIH Pub. No. 03-5370)
- Cannell, CF.; Fowler, FJ.; Marquis, K. *Vital and Health Statistics*. Washington, DC: U.S. Government Printing Office; 1968. The influence of interviewer and respondent psychological and behavioral variables on the reporting in household interviews. Series 2, No. 26
- Carlson ED. A case study in translation methodology using the Health-Promotion Lifestyle Profile II. *Public Health Nursing*. 2000; 17:61–70. [PubMed: 10675054]
- Census Bureau (U.S.). *Census Bureau Guideline: Language translation of data collection instruments and supporting materials*. 2004. Retrieved August 12, 2007 from <http://www.census.gov/cac/www/007585.html>
- Centers for Disease Control and Prevention (U.S.). *Tobacco use among U.S. racial/ethnic minority groups – African Americans, American Indians and Alaska Natives, Asian Americans and Pacific Islanders, and Hispanics: A report of the Surgeon General*. Atlanta: Georgia; 1998.
- DeMaio, TJ.; Rothgeb, JM. *Cognitive Interviewing Techniques: in the Lab and in the Field*. In: Schwarz, N.; Sudman, S., editors. *Answering questions*. California: Jossey-Bass; 1996. p. 177-198.
- DeVellis, RF. *Scale development: Theory and applications*. 2nd ed.. Thousand Oaks: Sage; 2003.
- Edwards, WS.; Levin, K.; Willis, G.; Lawrence, D.; Thompson, F. Comparing behavior coding and cognitive interviewing as assessment tools for survey questions on health topics. Paper presented at the meeting of the European Association of Survey Research; Barcelona: Jul. 2005

- European Social Survey. An outline of ESS translation strategies and procedures. 2002. Retrieved April 13, 2006 from http://naticent02.uuhost.uk.uu.net/ess_docs/R3/Methodology/r3_translation_guidelines.pdf
- Forsyth BH, Kudela MS, Levin K, Lawrence D, Willis GB. Methods for translating an English-language survey questionnaire on tobacco use into Mandarin, Cantonese, Korean, and Vietnamese. *Field Methods*. 2007; 19:264–283.
- Forsyth, BH.; Lessler, JT. Cognitive laboratory methods: A taxonomy. In: Biemer, PP.; Groves, RM.; Lyberg, LE.; Mathiowetz, NA.; Sudman, S., editors. *Measurement errors in surveys*. New York: Wiley; 1991. p. 393-418.
- Fowler, FJ. *Improving survey questions*. Thousand Oaks, CA: Sage; 1995.
- Fowler, FJ.; Cannell, CF. Using behavioral coding to identify problems with survey questions. In: Schwartz, N.; Sudman, S., editors. *Answering questions: Methodology for determining cognitive and communicative processes in survey research*. San Francisco: Jossey-Bass; 1996. p. 15-36.
- Goerman, P. *The Promised land? The lives and voices of Hispanic immigrants in the new South*. New York: Routledge; 2006.
- Harkness, JA.; Schoua-Glusberg, A. Questionnaires in translation. In: Harkness, JA., editor. *Cross-cultural survey equivalence*. Mannheim, Germany: ZUMA; 1998. p. 87-126.
- Harkness, JA.; Van de Vijver, FJR.; Mohler, P. *Cross-cultural survey methods*. Hoboken, N. J: Wiley; 2003.
- Hunter JC, Landreth A. Analyzing interviewer/respondent interactions while using a mobile computer-assisted personal interview device. *Field Methods*. 2006; 8:335–351.
- Ji LJ, Schwarz N, Nisbett RE. Culture, autobiographical memory, and behavioral frequency reports: Measurement issues in cross-cultural studies. *Personality and Social Psychology Bulletin*. 2000; 26:586–594.
- Johnson TP. Methods and frameworks for crosscultural measurement. *Medical Care*. 2006; 44:S17–S20. [PubMed: 17060823]
- Johnson, TP. Approaches to equivalence in cross-cultural and cross-national survey research. In: Harkness, JA., editor. *Cross-cultural survey equivalence*. Mannheim, Germany: ZUMA; 1988. p. 1-40.
- Kim SS, Ziedonis D, Chen KW. Tobacco use and dependence in Asian Americans: A review of the literature. *Nicotine and Tobacco Research*. 2007; 9:169–184. [PubMed: 17365748]
- Levy DT, Romano E, Mumford E. The relationship of smoking cessation to sociodemographic characteristics, smoking intensity, and tobacco control policies. *Nicotine & Tobacco Research*. 2005; 7:387–396. [PubMed: 16085506]
- Ma G, Tan Y, Toubbeh JI, Su X, Shive SE, Lan Y. Acculturation and smoking behavior in Asian-American populations. *Health Education Research*. 2004; 19:615–625. [PubMed: 15199009]
- Martinez G, Marín BV, Schoua-Glusberg A. Translating from English to Spanish: The 2002 National Survey of Family Growth. *Hispanic Journal of Behavioral Sciences*. 2006; 28:531–545.
- McKay RB, Breslow MJ, Sangster RL, Gabbard SM, Reynolds RW, Nakamoto JM, Tarnai J. Translating survey questionnaires: Lessons learned. *New Directions for Evaluation*. 1996; 70:93–105.
- Miller, K. Implications of socio-cultural factors in the question response process. In: Pruffer, P.; Rexroth, M.; Fowler, FJ., editors. *Questionnaire evaluation standards*. Mannheim, Germany: ZUMA; 2004. p. 172-188.
- Nápoles-Springer AM, Santoyo-Olsson J, O'Brien H, Stewart AL. Using cognitive interviews to develop surveys in diverse populations. *Medical Care*. 2006; 44(Suppl 3):S21–S30. [PubMed: 17060830]
- Nápoles-Springer AM, Stewart AL. Overview of qualitative methods in research with diverse populations: Making research reflect the population. *Medical Care*. 2006; 44(Suppl 3):S5–S9. [PubMed: 17060835]
- National Cancer Institute. *National cancer trends progress report - 2005 update*. National Cancer Institute, NIH, DHHS; Bethesda, MD: Dec. 2005 Retrieved May 6, 2007 from <http://progressreport.cancer.gov/>

- Pan, Y.; de la Puente, M. Census Bureau guidelines for the translation of data collection instruments and supporting materials: Documentation on how the guideline was developed. Washington, D.C.: Research Report Series #2005-06, Statistical Research Division, U.S. Bureau of the Census; 2005.
- Ponce NA, Lavarreda SA, Yen W, Brown ER, DiSogra C, Satter D. The California Health Interview Survey 2001: Translation of a major survey for California's multiethnic population. *Public Health Reports*. 2004; 119:388–395. [PubMed: 15219795]
- Pasick R, Stewart SL, Bird J, D'Onofrio CN. Quality of data in multiethnic health surveys. *Public Health Reports*. 2001; 116(Suppl):223–243. [PubMed: 11889288]
- Quittner AL, Sweeny S, Watrous M, Munzenberger P, Bearss K, Nitza AG, Fisher LA, Henry B. Translation and linguistic validation of a disease-specific quality of life measure for Cystic Fibrosis. *Journal of Pediatric Psychology*. 2000; 25:403–414. [PubMed: 10980045]
- Reeve, BB. Item response theory. *Polling America: An encyclopedia of public opinion*. Best, S.; Radcliff, B., editors. Westport, CT: Greenwood; 2005. 2005. p. 400-410.
- Rivara FP, Ebel BE, Garrison MM, Christakis DA, Wiehe SE, Levy DT. Prevention of smoking-related deaths in the United States. *American Journal of Preventive Medicine*. 2004; 27:118–125. [PubMed: 15261898]
- Rogler LH. Methodological sources of cultural insensitivity in mental health research. *American Psychologist*. 1999; 54:424–433. [PubMed: 10392472]
- Schmidt S, Bullinger M. Current issues in cross-cultural quality of life instrument development. *Archives of Phys Med Rehabil*. 2003; 84:S29–S34.
- Singelis TM, Yamada AM, Barrio C, Laney JH, Her P, Ruiz-Anaya AR, Lennertz ST. Metric equivalence of the Bidimensional Acculturation Scale, the Satisfaction With Life Scale, and the Self-Constraint Scale across Spanish and English language versions. *Hispanic Journal of Behavioral Sciences*. 2006; 28:231–244.
- Stewart AL, Nápoles-Springer A. Health-related quality-of-life assessments in diverse population groups in the United States. *Medical Care*. 2000; 38(Supplement II):102–124.
- Tourangeau, R.; Rips, LJ.; Rasinski, K. *The psychology of survey response*. Cambridge: Cambridge University Press; 2000.
- Warnecke RB, Johnson TP, Chavez N, Sudman S, O'Rourke DP, Lacey L, Horm J. Improving question wording in surveys of culturally diverse populations. *Annals of Epidemiology*. 1997; 7:334–342. [PubMed: 9250628]
- Werner, O.; Campbell, DT. Translating, working through interpreters and the problem of decentering. In: Cohen, RN.; Cohen, R., editors. *A handbook of method in cultural anthropology*. New York: American Museum of Natural History; 1970. p. 398-420.
- Willis, G. *Cognitive interviewing: A tool for improving questionnaire design*. Thousand Oaks, CA: Sage; 2005.
- Yu DSF, Lee DTF, Woo J. Issues and challenges of instrument translation. *Western Journal of Nursing Research*. 2004; 26:307–320. [PubMed: 15068554]
- Zahnd, E.; Tam, T.; Lordi, N.; Willis, G.; Edwards, S.; Fry, S.; Grant, D. Cross-cultural behavior coding: Using the 2003 California Health Interview Survey (CHIS) to assess cultural/language data quality. Barcelona: Paper presented at the meeting of the European Association for Survey Research; Jul. 2005

Table 1

Topics within the 2003 Tobacco Use Special Cessation Supplement to the Current Population Survey (TUSCS-CPS)

- Cigarette smoking prevalence
- Smoking history
- Current and past cigarette consumption
- Cigarette smoking quit attempts and intentions to quit
- Medical and dental advice to quit smoking
- Cigar, pipe, chewing tobacco, and snuff use
- Workplace smoking policies
- Smoking rules in the home
- Attitudes toward smoking in public places
- Opinions about the degree of youth access to tobacco in the community
- Level of nicotine addiction
- The nature of quitting methods, products, and treatment information
- Quitting history and methods used for cigarettes and for other forms of tobacco
- Cost of last pack/carton of cigarettes purchased

Table 2

Steps in translating the TUSCS-CPS into Spanish and Asian languages and empirically evaluating the translations¹

Step:	Description of Activities	Staff involved
Step 1: Translation	<ul style="list-style-type: none"> Created draft target language questionnaires by translating source questionnaire Documented translation challenges and decisions on how to address them. 	<ul style="list-style-type: none"> Multilingual teams consisting of professional translators (e.g., English-Mandarin-Cantonese; English-Spanish)
Step 2: Review	<ul style="list-style-type: none"> Reviewed translated questionnaires and translators' documentation. Identified specific recommendations for changes to translations. 	<ul style="list-style-type: none"> Survey Language Consultants (SLCs) having extensive language skills, knowledge of cultural group, and familiarity with research objectives of survey
Step 3: Initial Adjudication	<ul style="list-style-type: none"> Reviewed suggestions for revision and make decisions, resulting in pretest-ready target language questionnaires. 	<ul style="list-style-type: none"> Bilingual professional language reviewers (University consultants); contract staff; SLCs
Step 4: Cognitive interview pretest	<ul style="list-style-type: none"> Conducted intensive, individual interviews: 9 each in Spanish, Chinese (5 Mandarin and 4 Cantonese), and Korean; 13 in Vietnamese². Analyzed results, and developed recommendations for changes based on testing results. 	<ul style="list-style-type: none"> SLCs and bilingual interviewers trained in cognitive interviewing techniques Language reviewers and English-speaking project staff
Step 5: Second Adjudication	<ul style="list-style-type: none"> Determined which recommendations to accept and what revisions to make to the target-language questionnaires. 	<ul style="list-style-type: none"> Project Director, advised by SLCs and English-speaking project staff
Step 6: Pilot Test	<ul style="list-style-type: none"> Questionnaires were empirically evaluated through 418 telephone-based interviews using Behavior Coding 	<ul style="list-style-type: none"> Contractor project and field staff Bilingual behavior coders and multilingual SLC
Step 7: Final Adjudication	<ul style="list-style-type: none"> Made final changes to questionnaire content for each translation based on pilot test results 	<ul style="list-style-type: none"> Project director, advised by English speaking project staff
Step 8: Documentation	<ul style="list-style-type: none"> Prepared website; Developed reports announcing availability of translations and providing details on how they were developed 	<ul style="list-style-type: none"> Authors of this manuscript

¹For more details, see Forsyth, et al. (2007).

²The first round of interviewing produced questionable results, seemingly due to SLC and interviewer inexperience, so the investigators conducted a second round using a different interviewing staff.

Table 3

Demographics characteristics of Pilot Test respondents, by language group

Demographic characteristic	English (n=125)	Spanish (n=51)	Mandarin (n=57)	Cantonese (n=67)	Korean (n=66)	Vietnamese (n=52)
Age*						
18 to 30	20.0% (25)	5.9% (3)	5.3% (3)	3.0% (2)	10.6% (7)	1.9% (1)
31 to 50	40.8% (51)	29.4% (15)	43.8% (25)	38.8% (26)	34.8% (23)	53.9% (28)
51 or older	25.6% (32)	58.8% (30)	50.9% (29)	58.2% (39)	53.1% (35)	44.2% (23)
Missing	13.6% (17)	5.9% (3)	0.0% (0)	0.0% (0)	1.5% (1)	0.0% (0)
Gender*						
Male	58.4% (73)	56.9% (29)	93.0% (53)	92.5% (62)	89.4% (59)	94.2% (49)
Female	34.4% (43)	39.2% (20)	7.0% (4)	7.5% (5)	10.6% (7)	5.8% (3)
Missing	7.2% (9)	3.9% (2)	0.0% (0)	0.0% (0)	0.0% (0)	0.0% (0)
Smoker type						
Every day	73.6% (92)	70.6% (36)	75.4% (43)	71.6% (48)	81.8% (56)	82.7% (43)
Some days	11.2% (14)	17.7% (9)	15.8% (9)	13.4% (9)	9.1% (6)	11.5% (6)
Former	15.2% (19)	11.8% (6)	8.8% (5)	14.9% (10)	9.1% (6)	5.8% (3)
Years in the U.S.*						
10 years or less	8.8% (11)	3.9% (2)	33.3% (19)	17.9% (12)	13.6% (9)	25.0% (13)
More than 10 years	27.2% (34)	70.6% (36)	66.7% (38)	82.1% (55)	84.9% (56)	73.1% (38)
Entire life	64.0% (80)	25.5% (13)	0.0% (0)	0.0% (0)	1.5% (1)	0.0% (0)
Missing	0.0% (0)	0.0% (0)	0.0% (0)	0.0% (0)	0.0% (0)	1.9% (1)
Education level*						
Less than high school	12.8% (16)	50.9% (26)	22.8% (13)	41.8% (28)	4.6% (3)	27.0% (14)
High school graduate	32.8% (41)	27.5% (14)	26.3% (15)	40.3% (27)	25.8% (17)	38.5% (20)
More than high school	54.4% (68)	21.6% (11)	50.9% (29)	17.9% (12)	65.2% (43)	32.6% (17)
Missing	0.0% (0)	0.0% (0)	0.0% (0)	0.0% (0)	4.5% (3)	1.9% (1)

* Distribution differs significantly across language groups at $p < .05$ level, by Chi-square test. All statistical tests were restricted to substantive (non-missing) categories, and categories were pooled when necessary to obtain cell sizes adequate for the conduct of statistical testing.

Table 4

Behavior codes used within Pilot Test to assess interviewer and respondent performance.

<i>Interviewer-based codes:</i>	
1	Question read correctly
2	Interviewer <u>Skips</u> question reading
3	<u>Reading Error</u> (mis-read): Interviewer fails to read the question as scripted
<i>Respondent-based codes:</i>	
1	Respondent provides <u>Codeable</u> (appropriate) response *
2	Respondent <u>Interrupts</u> (with an answer to the question, before the interviewer is done reading it)
3	Respondent requests that the interviewer <u>repeat</u> the question or request <u>Clarification</u> ("What do you mean by ...")
4	Respondent answers but provides an <u>Uncodeable</u> response ("After I get up I wait an hour or two before my first cigarette")

* Note that a codeable response is not one that is necessary correct – only that it meets the requirements of the survey data collection process.