Evaluation of a Multi-Step Survey Translation Process

Gordon B. Willis,[1] Pat Dean Brick,[2] Alicia Norberg,[2] Debra S. Stark,[2] Martha Stapleton Kudela,[2]
Barbara H. Forsyth,[3] Kerry Levin[2]
David Berrigan,[1] Frances E. Thompson,[1] Deirdre Lawrence[1]

The survey world is currently witnessing a surge of interest in multilingual research. In the U.S, this development is propelled by an increasingly diverse population, whereas, in Europe a similar interest is, at least in part, the byproduct of the growth of the European Union and its movement towards unification, standardization, and harmonization. U.S. surveys that track the wellbeing of the population are now expanding from one, or a few languages, to a large number. For example, in 2001 the California Health Interview Survey was conducted in 7 languages: English, Cantonese, Mandarin, Khmer, Vietnamese, Korean, and Spanish.

As a requisite condition for harmonization of versions, the survey literature contains a multitude of conceptualizations of the construct of comparability and equivalence. A common theme that runs through the literature is that 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, et al. (2006) found that the phrase "medical tests and procedures" was particularly problematic for Latinos because it failed to bring to mind particular events, and that the addition of examples (blood test, x-ray, cancer screening tests) to all versions would likely enhance comparability.

An ultimate translation goal is to "ask the same question" in more than one language. This goal is conceptually simple and reflects an operational approach and mindset. Further, it

[1] National Cancer Institute
[2] Westat
[3] University of Maryland Center for the Advanced Study of Language

expresses the desire that the estimates derived from the survey data demonstrate cross-cultural comparability across all important dimensions (Miller, 2004; Nápoles-Springer & Stewart, 2006; Pan & de la Puente, 2005; Schmidt & Bullinger, 2003; Singelis, Yamada, Barrio, et al., 2006; Yu, Lee, & Woo, 2004). Despite its apparent simplicity, there are a number of theoretical perspectives embedded in this goal. One interpretation of "asking the same question" is articulated by Census Bureau (2004): that the translation be reliable, complete, accurate, and culturally appropriate, and demonstrate semantic, conceptual, and normative equivalence.

For several decades, translation efforts have emphasized back translation (Brislin, 1970). In back translation, one translator produces a translation into the target language. The target-language version is then translated back to the source language by a different translator. The two versions in the source language document are compared and the quality of the target language document evaluated, based on the degree of literal correspondence between the two source language versions. The problems inherent in such a practice are increasingly being noted by survey researchers, however (see Ponce, et al, 2004 for more detail), and recent developments in survey translation practice are deliberate departures from the back-translation method. These methods largely endeavor to avoid an overly literal, word-for-word rendition that may fail to communicate the measurement goals of the question similarly across version. To circumvent such deleterious effects, newer practices offers a novel approach to both the translation exercise itself, and to the procedures used to evaluation the translation – including the addition of empirical methods of survey pretesting that were previously reserved mainly for single language questionnaires (Carlson, 2000; Census Bureau, 2004; European Social Survey, 2002; Forsyth, Kudela, Levin, et al., 2007; Harkness, et al., 2003; Harkness & Schoua-Glusberg, 1998;).

Specifically, cross-disciplinary research has assisted in the development of multi-step, committee-based translation methodologies. Harkness et al. (1998; 2003) describe the framework for a five-step process that involves multiple levels of review and reconciliation: Translate, Review, Adjudicate, Pretest, and Document (TRAPD). The TRAPD framework sets out a number of theoretical and conceptual considerations that lead to a multifaceted view emphasizing both linguistic and socio-cultural elements. In the U.S., the Census Bureau (2004) has also outlined a five-step process that similarly employs a team approach: Prepare, Translate, Pretest, Revise, and Document (PTPRD). The Census guide also addresses specific operational issues at a relatively high level of detail, such that even a novice practitioner may be able to execute a methodologically sound translation project.

At present, there is a paucity of literature that presents actual data on the contribution of the TRAPD process (or, the PTRPD process) to the overall quality of the translated survey instrument. A variety of survey research organizations have adapted and operationalized a TRAPD-inspired methodology – and the details of each application likely reflects the research needs and operational constraints of that organization. This chapter describes an initial process evaluation of Westat's adaptation of the TRAPD methodology. We present the results of five studies for which our implementation of the TRAPD framework was used as the translation and evaluation methodology. We report the types of questionnaire design issues that were identified, and potentially remediated, at three critical steps in the TRAPD process. We consider both quantitative and qualitative aspects of these data, in order to develop conclusions concerning the effectiveness of the methodology in the translation process.

The Westat TRAPD translation process. The current TRAPD application generally entails a five-step process that affords opportunities at each step for evaluating and revising translated materials. The revisions that derive from a given step are used as inputs for the following step, with a goal of continuous improvement (see Table 1).

Table 1. Description of the TRAPD survey translation process used within the current study.

1. Translation	Develop a survey translation using a committee approach
2. Review*	Expert review of the translation(s) to identify problems and additional translation options
3. Adjudication*	Make decisions to reconcile options from preceding two steps
4. Pretesting: Cognitive interviewing* Behavior coding of survey pretest	Intensively interview language-appropriate respondents to identify difficulties in understanding and answering the questions, and to identify translation issues that impede comparability Conduct a field test of the survey translation and use observational methods to identify potential problems with translated versions
5. Documentation	At each step, compile qualitative and quantitative data that may be reviewed and further coded for purposes of report-writing, quality-control assessment, and evaluation research

*Step that was evaluated in the current investigation

The Westat application of TRAPD is usually executed in the following manner:

Step 1: Translation. The English language source questionnaire is translated by either an in-house team of translators, by a translation vendor, or by a consultant (individual or team). Instructions to the translators can vary by project. In some projects, translators are instructed to

document problems and issues that emerged during the translation process. For these projects, records are produced that chronicle the issues that the initial translators felt were problematic. On other projects, the initial translators resolve the issues among themselves without generating a record of issues encountered.

Steps 2 and 3: Review and Adjudication. At both of these Steps, multilingual staff review the initial translations and the suggested modifications. Review and Adjudication are based on multilingual staff's familiarity with cultural norms and knowledge of survey measurement goals, questionnaire design and survey methods-- and therefore extends beyond the level of such knowledge generally represented by the initial translating team. In turn, Reviewers and Adjudicators are asked to justify their revisions, thus generating a record of problems encountered, and of anticipated defects in the translated items. In some cases, the Reviewer/Adjudicator is given a template to work from; and in others, they recorded their comments and suggestions using their choice of reporting format. The evaluated questionnaire translations are first modified as part of the Adjudication process, in which the initial Translated version, and the annotated version produced through the Review process, are reconciled based on the Adjudicator's assessment of the most workable approach. For the current studies, both Review and Adjudication processes produced extensive qualitative data concerning identified problems with the items evaluated at those steps, which allowed a systematic analysis of the products of these Steps.

Step 4: Pretesting: Table 1 illustrates a comprehensive pretesting approach in which cognitive interviewing (described in depth by Willis, 2005) and behavior coding (Fowler & Cannell, 1996) are applied in turn. Because behavior coding was used in only one of the Studies

described here, the current paper is limited to evaluation of the outcomes of cognitive interviewing, as it constituted the Pretesting step.

Step 5: Documentation. Finally, documentation of the potential problems that were identified were available from the Review, Adjudication, and Cognitive Testing steps. These three steps were therefore the focus of the current evaluation.

Applying the TRAPD process in five studies involving questionnaire translation

We applied variants of the five-step process in Table 1 to develop questionnaire translations in five studies. The studies involved different target languages (across all Studies: Cantonese and Mandarin Chinese, Korean, Spanish and Vietnamese) and different survey content (tobacco use, diet, acculturation to U.S. society, and physical activity). Data from the Review, Adjudication, and Cognitive Interviewing (Pretesting) Steps were analyzed to assess the problems identified at each of these Steps, for each Study. We first describe each study briefly:

Study 1: The Tobacco Use Supplement to the Current Population Survey (TUS-CPS). The Tobacco Supplement (TUS) is a periodic surveillance survey (administered both face-face and over the telephone) used as a key source of data on smoking and other tobacco use in U.S. National and State-level household populations. The initial TUS translation was produced by both translation vendors and in-house Westat translation staff. These translators worked independently to produce four target-language questionnaires (Chinese, Vietnamese, Korean, and Spanish). Review of these translations was conducted by survey language consultants (SLCs) expert in the target languages. The Review step was designed to be in-depth and intensive, with detailed suggestions for revisions as the outcome. The SLCs were given a detailed template to assist them in structuring and documenting their comments. Adjudication was conducted by

subject-matter experts with both translation experience and a strong survey methods background. Cognitive interviewing involved 41 total subjects, and is described in more detail by Willis, Lawrence, Hartman, Stapleton Kudela, Levin, and Forsyth (in press).

Study 2. The National Health Interview Survey. The NHIS is a periodic, nationwide household interview survey conducted by the National Center for Health Statistics in approximately 40,000 U.S. households. The Cancer Control Module collects data on a wide array of cancer risk factors and was translated from English into Spanish. Professional translators from the Library of Congress produced an initial Spanish translation using previously translated questions as reference material. The Review step was conducted by a multi-agency bilingual review team with content, language, and methodological expertise. Adjudication was conducted by an independent Adjudicator who made the final decisions about revisions prior to cognitive testing. Bilingual survey methodologists conducted two rounds of cognitive interviews; the instruments were revised between the two rounds of cognitive testing, and after the second round.

Study 3. California Health Interview Survey. The CHIS is a random digit dialing survey of about 65,000 completed telephone interviews that measure topics such as health insurance coverage and access to health services within the State of California. In various years, CHIS has been conducted in as many as seven languages; this particular project describes the translation process for the Chinese version. A professional translation vendor produced the Chinese-language questionnaire that was delivered to Westat for review. Review was conducted by external experts hired and trained by Westat. Adjudication was also conducted by an external expert who made the final decisions about changes to the instrument. Subsequently, two rounds

of cognitive testing were conducted with Chinese-speaking respondents of varying ages, income levels, countries of origin, and educations by bilingual staff with varying degrees of survey research experience. The questionnaire was revised between the two rounds of cognitive testing, and following the second round.

Studies 4 and 5: Acculturation and Physical Activity**.** Westat simultaneously tested two independent sets of questions for NCI, centered on (ACC) acculturation of the various Hispanic groups to U.S. society; and (PA) physical activity within everyday life. Although these questions were Translated, Reviewed, Adjudicated, and Pretested together, we treat these as separate studies for current purposes, given the very different natures of the two topics and potential for fundamentally disparate varieties of problems to emerge. Translation was initially done by a single native Spanish speaker of South American origin who had prior survey translation experience. Review was carried out by two independent bilingual specialists with survey research experience and different Spanish-language backgrounds. Adjudication was by an experienced, independent expert who made the final decisions regarding all revisions prior to pretesting. Cognitive interviews were conducted by bilingual survey methodologists who also had extensive translation experience. Tested respondents varied by the length of domicile in the United States: One respondent group consisted of Latinos of long tenure who preferred to conduct the interview in English; the other group was composed of monolingual Spanish speakers.

Method

Coding of outcomes of Review, Adjudication and Cognitive Interview Steps**.** Because our objective was to document and quantify the outcomes of each of these steps, in as

quantitative a manner as feasible, we developed two independent coding systems that could be applied to the text-based results of each Step. The written comments that were produced for each question, at each Step, and for each Study, were coded by Westat staff using two coding systems that were derived from earlier evaluation of pretesting outcomes. One system, labeled the TCG (Translation, Cultural adaptation, Generic problem system) had been used previously by Willis, et al. (in press) to characterize the outcomes of multi-lingual, cross-cultural cognitive interviews. The system is based on the finding that such results generally have fallen into three relatively distinct categories (see Appendix 1 for a complete description):

(a) T: Those that are linguistic in nature, and due to defects in the Translation process (e.g., a term that is mistranslated)

(b) C: Those that transcend language in the structural sense, but represent problems of Cultural adaptation in which questions do not function appropriately in one or more groups due to specific features of that group (e.g., problems of worldview, or structural differences between societies that create logical defects in the item)

(c) G: Generic problems of question design that appear to affect all tested groups, and are not culturally specific (e.g., general difficulty in recall of long-ago events)

Because we were unsure whether the TCG system would be adequate, we also applied a separate coding system, based on a prior system – the Q-BANK (Miller, 2006) designed to classify potential problems with survey items. Because the Q-BANK was designed for monolingual questionnaires, we developed a revision of the system, labeled the MQ-BANK (i.e., Modified Q-BANK), in order to incorporate codes appropriate for translated questions (see Appendix 1). In particular, we developed a specific code (Semantics/Syntax) that was intended

to specifically identify any problems with the structural characteristics of items that were specific to translation.

Coding procedures. Study documentation provided detailed qualitative data that we coded, according to the two coding systems described above. As such we were able to use as raw material the written justification for recommendations (i.e., the anticipated or observed problems with each item) at the Review, Adjudication, and Cognitive testing Steps. Table 2 summarizes the analytic approach and contains information about the process and products associated with each step.

Table 2. Overview of 3-step analytic approach

Analytic Step	Process	Purpose
Review documentation and detect problems	Identify problems based on evidence in each document for Review, Adjudication, and Cognitive Testing Steps.	To identify problem items in preparation for coding of problem types.
Code identified problems by problem type	Applied TCG and MQ-BANK coding systems to describe types of problems found at each Step.	To characterize precise issues found in each Study, at each Step.
Tabulate problem Codes and make comparisons across Steps and across Studies	Compare frequency and type of problems across Steps and across Study, collapsing over tested questions.	To explore the similarities and differences of results, across Step and Study.

The heart of the analytic approach in Table 2 is identifying problems at each Step of the process (that is, the RAP Steps of the total TRAPD process) by systematically coding the types of problems uncovered at each step. Note that we have in all cases collapsed across item; rather

than retaining information about particular survey questions, we chose to quantify at the level of Code (under both coding system), to reveal the distribution of these codes across Step and Study.

As depicted in analysis Step 2 of Table 2, coders identified problems whenever an item description indicated potential response error. The coders were four Westat survey methodologists who had at least some previous coding experience. Coder training consisted of first reviewing a training document as a group, to learn the coding systems; and then distributing the to-be-coded documents from the five Studies for independent coding under both systems. A set of coding rules was applied such that (a) a questionnaire item could have more than one problem, such that one item could receive more than one code; and (b) a problem that recurred often, but was essentially a single issue, was coded only once. For example, an erroneous translation of a term that occurred repeatedly throughout the questionnaire was coded as such once, rather than every time it occurred. Once coding was completed, the codes were entered into a spreadsheet where they were aggregated and cross-tabulated.

Results

Problems identified at each Step: Review, Adjudication, Cognitive Testing. The first issue addressed in the analysis concerned the nature of coded problems in translated items that were identified by each of the three evaluation Steps (collapsed across individual codes). Table 3 summarized the distribution of these problems for each Study, broken down by Step. Overall, each of the Steps was generally successful in identifying problems, although the relative frequencies of these problems varied across studies. In all Studies, either the Review or Adjudication Step identified the highest relative frequency of problems, and the Cognitive Interviewing Step was either second or third, quantitatively (for example, in the NHIS, the

respective percentages of problems identified at Review, Adjudication, and Cognitive testing were fairly even, at 38.5%, 25.3%, and 36.1%) . Given that the Cognitive Interviews were done last in the series, it may not be surprising that they generally produced a smaller number of problems.

Table 3. Distribution of coded problems as a function of Study and Step.

	TUS		NHIS		CHIS		ACCULTURA-TION		PHYSICAL ACTIVITY	
Step	n	%	n	%	n	%	n	%	n	%
Review	188	39.00	32	38.55	33	55.00	13	32.50	22	53.66
Adjudication	228	47.30	21	25.30	9	15.00	17	42.50	14	34.15
Cognitive Interview	66	13.69	30	36.14	18	30.00	10	25.00	5	12.20
Total	482	100.00	83	100.00	60	100.00	40	100.00	41	100.00

Problems identified by TCG and MQ-BANK coding systems. Next, we focused on the results of individual codes, as applied to the problems identified across all Studies and Steps. We initially cross-tabulated the two coding systems, collapsing across all other factors, to assess any discernible pattern of overlap between these systems. The TCG system revealed most problems (608/706, or 86.1%) to be located in the T (Translation) category. Generic problems, or those judged to persist across groups, were the next most frequent (73/706, or 10.3%). Somewhat surprisingly, the Cultural problems that are of particular interest to many cross-cultural researchers were fairly infrequent (25/706, or 3.5%).

Under the alternate, MQ-BANK coding system, the bulk of problems were located in the code labeled Semantics/Syntax, which was added to the Q-BANK system mainly to accommodate translation difficulties associated with multiple language versions (54.1%). Other

prominent problems were typographical or formatting problems (13.2%), and Ambiguous concepts (6.1%). Overall, the two coding systems converged, in identifying mainly problems that involved translation defects.

Assignment of Codes Across Study. At a more detailed level of analysis, we next ascertained the distribution of codes (under both TCG and MQ-BANK coding systems) at the level of the individual Study (Tables 4 and 5), to determine the degree to which problem types were common, as opposed to specific to Study. Under the TGC coding system, Translation was by far the most frequently applied code, in all five Studies, and the relative frequencies of Cultural and Generic problems varied across Study (e.g., in the CHIS, only 1 Generic problem, but 9 Cultural-adaptation problems, were identified; whereas in the Physical Activity Study, the pattern was opposite). As a convergent result, for the more detailed MQ-BANK code system, Translation-related codes (in particular, those related to Semantics/Syntax) again dominated, in each study. Overall, translation problems clearly outweighed any other form of identified problem, in all evaluated Studies.

Table 4. Frequency and Percentage of TCG Problem Type, by Study.

	TUS		NHIS		CHIS		ACCULTURATION		PHYSICAL ACTIVITY	
TGC Problem Code	n	%	n	%	n	%	n	%	n	%
Translation	444	92.1	58	69.9	50	83.3	26	65.0	30	73.2
Culture Related	5	1.0	5	6.0	9	15.0	6	15.0	0	0.0
Generic Design	33	6.9	20	24.1	1	1.7	8	20.0	11	26.8
TOTAL	482	100.0	83	100.0	60	100.0	40	100.0	41	100.0

Table 5. Frequency and Percentage of MQ-BANK Problem Type, by Study.

	TUS		NHIS		CHIS		ACCULTURATION		PHYSICAL ACTIVITY	
MQ-BANK Problem Code	n	%	n	%	n	%	n	%	n	%
Interviewer Difficulty	1	.21	0	0.0	0	0.0	0	0.0	0	0.0
Term problem	10	2.07	4	4.8	4	6.7	0	0.0	0	0.0
Ambiguity	14	2.90	14	16.9	10	16.7	3	7.5	2	4.9
Overly complex	2	.41	3	3.6	0	0.0	4	10.0	0	0.0
Logical problem/assumption	5	1.04	0	0.0	0	0.0	0	0.0	0	0.0
Recall problem	4	.83	1	1.2	0	0.0	0	0.0	4	9.8
Bias/Sensitivity	2	.41	0	0.0	0	0.0	0	0.0	0	0.0
Questionnaire problem	3	.62	2	2.4	0	0.0	0	0.0	0	0.0
Response Option problem	5	1.04	0	0. 0	0	0.0	4	10.0	0	0.0
Can't be Translated	1	.21	3	3.6	8	13.3	4	10.0	1	2.4
Other	84	17.43	4	4.8	4	6.7	7	17.5	13	31.7
Typographical problem	71	14.73	14	16.9	4	6.7	3	7.5	1	2.4
Semantic or Syntactical problem in translation	280	58.09	38	45.8	30	50.0	15	37.5	20	48.8
TOTAL	482	100.0	83	100.0	60	100.0	40	100.0	41	100.0

Problem distribution across Step. Finally, in order to examine the extent to which the evaluated Steps detected similar types of problems, Tables 6 and 7 reveal the manner in which specific problem codes distributed across Step (Review, Adjudication, and Cognitive Interviewing). Under the TCG system, Review and Adjudication Steps focused almost exclusively on Translation problems (as they are nominally designed to do). Interestingly, Cognitive Interviewing, as an evaluative Step, was much more balanced in this regard; that procedure overall located more Generic questionnaire problems than either Translation or Cultural problems, yet the latter two categories were still fairly frequently represented. Overall, over 80 percent of the total Generic and Cultural problems were identified through cognitive testing. As such, Cognitive Interviewing appears to provide the widest "net" of any of the evaluated Steps, in terms of variety of problems identified. Similarly, for the MQ-BANK system, Review and Adjudication Steps again concentrated on structural/linguistic issues relating to Translation, whereas Cognitive Interviewing again focused largely on general problems involving ambiguity of terms and whole questions.

Table 6. Distribution of TCG problem codes identified at each evaluation Step.

	Culture Related		Generic Design		Translation	
	n	percent	n	percent	n	percent
Review	1	4.00	5	6.85	282	46.38
Adjudication	2	8.00	7	9.59	280	46.05
Cognitive Interview	22	88.00	61	83.56	46	7.57
Total	25	100.00	73	100.00	608	100.00

Table 7. Distribution of MQ-BANK problem codes identified at each evaluation step.

MQ-BANK Problem Code	Review		Adjudication		Cognitive Interview	
	n	%	n	%	n	%
Interviewer Difficulty	0	0.00	0	0.00	1	0.78
Term problem	0	0.00	3	1.04	15	11.63
Ambiguity	3	1.04	0	0.00	40	31.01
Overly complex	1	0.35	4	1.38	4	3.10
Logical problem/assumption	0	0.00	0	0.00	5	3.88
Recall problem	0	0.00	0	0.00	10	7.75
Bias/Sensitivity	0	0.00	2	0.69	0	0.00
Questionnaire problem	0	0.00	0	0.00	5	3.88
Response Option problem	1	0.35	0	0.00	8	6.20
Cannot be Translated	9	3.13	1	0.35	7	5.43
Other	21	7.29	79	27.34	12	9.30
Typographical problem	12	4.17	79	27.34	2	1.55
Semantic or Syntactical problem in translation	241	83.68	121	41.87	20	15.50
TOTAL	288	100.00	289	100.00	129	100.00

Discussion

Based on these patterns of results, we make several conclusions and recommendations for translation, in the context of appropriate caveats and suggestions for further research that will clarify unresolved issues. First, based on the finding that multiple Steps each identified non-trivial levels of problems, we believe it to be useful to rely on a multi-step approach. It may be best to further incorporate a behavior coding step, as Willis et al. (in press), as this allows the researchers to evaluate the operation of each questionnaire version in a field environment. Second, we conclude that, despite the prior efforts of the developers, Generic problems of question design are common, and difficult to eradicate. Certainly the process of effective translation may not resolve these problems, as they are simply carried through from the source version to each translation. However, the mechanisms used to evaluate the translations – and especially cognitive interviewing – appear to be useful in identifying these issues.

Significantly, problems of cultural adaptation – where between-group differences in worldview or social behavior transcend language translation, and produce non-comparability of survey items – appeared to be less frequent than either pure translation problems or general issues. Of course, one can reasonably argue that such problems do exist (as we did find across most of the Studies), and that they pose particularly severe threats to data quality, as they very likely produce bias in cross-cultural comparisons (unlike Generic problems, which would be expected to simply add "noise" equally to all contrasted groups). As such, it is reassuring that these problems do emerge, again largely as a function of cognitive interviewing, which may be especially proficient at bringing otherwise hidden problems to the surface.

Finally, and in summary, we note that the evaluation Steps, within a TRAPD framework, differ in what they accomplish. Cognitive interviewing may be the least productive Step quantitatively, in that it identified a smaller absolute number of problems in most cases than either Review or Adjudication (and this finding is consistent with other research that quantifies the results of pretesting, such as Presser and Blair, 1994; Willis, 2005) From a qualitative point of view, however, Cognitive Interviewing may be vital in focusing on what DeMaio and Rothgeb (1996) have labeled "Silent Misinterpretations" that are only elucidated through the process of verbal probing of actual test respondents.

Caveats

The conclusions above must be tempered by several limitations of the current study. First, we note that the current focus was process evaluation, as opposed to evaluation of verifiable outcome measures. Therefore, we possessed no absolute criterion measure of either item quality or response error, by which to determine whether the problems that were recorded and then coded as part of each of the five Studies were real ones, as opposed to artifacts of the research procedures. However, we note that the types of problems identified are very similar to those described in previous cross-cultural studies (e.g., Miller, Willis, Eason, Moses, & Canfield, 2005; Miller, et al., in press), and therefore appear to be fairly ubiquitous in such studies.

A second limitation involves scope: Despite the considerable amount of effort that we expended in aggregating and then coding results across five investigations, the current study is limited to one country, and to a single organization. It is therefore unclear whether the results extend to the broader environment consisting of multiple locations (or countries), and across multiple questionnaire-design and evaluation organizations. We were also admittedly

constrained in scope, in the sense that our evaluation of the TRAPD translation and evaluation model focused on only three of the key steps (Review, Adjudication, and Pretesting). In some ways this may be appropriate, to the extent that the first and final steps (Translation and Documentation) are viewed as procedural steps, rather than evaluative ones, and do not generally give rise to data of the type that is analyzable using the current approach.

Finally, even to the extent to which we have taken a reasonable approach in evaluating a TRAPD model, it remains the case that we have evaluated one variant of that model, and have not taken the clear next step of contrasting the results of implementation of this model with those obtained from some alternative approach. As such, we recommend further research that explicitly compares alternatives, such as back-translation, or PTPRD, to the TRAPD framework, in order to determine which appears to be most effective and efficient.

In conclusion, we suggest that a multi-step model such as TRAPD appeared to be effective in identifying potential problems in the translation process, across a range of studies. This approach may represent an effective broadening of expert review and pretesting procedures in a way that will lead to enhancements of data quality across multi-lingual, multi-cultural, and cross-national studies.

Appendix 1

CODING DEFINITIONS

LEVEL 1 CODES [T/G/C]

CODE	DEFINTION	EXAMPLE
Translation problem	**Translated item altered intent of original question (including using the wrong level of formality).** **A wording problem is assumed to be translation-related unless the document (review, adjudication, cognitive interview report, etc.) specifically states there is a problem with the English-language version.**	A few [Chinese] characters were added/changed to make the sentence flow better. The translation of this item appears to be incorrect. The English version reads, "Have you ever switched from a stronger cigarette to a lighter cigarette for at least 6 months?" The Chinese translates, "Since you have switched from regular to light cigarettes, has it been more than half a year?"
Generic-design problem	**Difficulties understanding or answering a test item, apparently independent of culture or language.**	The English version is slightly awkward to begin with… Some respondents answered that less harm to their health and to help quitting smoking were equally important reasons for switching from full flavor to lighter cigarettes. However, the response options do not include "both" as a choice.
Culture-related problem	**Intended meaning difficult to convey due to culture-related difference in underlying constructs or culture-related conventions.**	A question about having switched from stronger to light cigarettes may pose a problem for respondents who started smoking a Korean brand of cigarettes (which does not list tar and nicotine amounts on the package) and later switched to an American brand.

LEVEL 2 CODES [MODIFIED Q-BANK]

CODE	DEFINTION	EXAMPLE
PERCEPTION		
Interviewer Difficulty	**Question is difficult to administer from interviewer perspective** • Cannot be read in a standardized manner • Unclear what parts of the question should be read • Interviewer cannot administer without obtaining information not previously collected	The following questions require the interviewer to make decisions regarding which parts of the question to read aloud to the respondent: What was the date of your most recent injury/poisoning? In the last year, was anyone in your household poisoned? Read when necessary: Do not include sun or food poisoning or reaction to poison ivy. (I'er read the statement before hearing the R's answer.)
COMPREHENSION		
Problematic Terms	**Contains inappropriate or unfamiliar terms** • Unknown terms • Terms used out of context or inappropriately • Overly technical language/jargon	Have you ever used a dental sealant? (respondents did not know term 'dental sealant') How often do you use a dosimeter? (respondents did not know term 'dosimeter')
Ambiguous Concepts	**Contains vague or ambiguous concepts** • Ambiguous or vague concepts or terms • Missing reference period • Concepts with multiple interpretations - answer could vary depending upon which interpretation was taken • Instructions or definitions are missing or inadequate	How many hours a week do you work near a large electrical machine? (A respondent constantly passed by a machine for a few seconds, and was unsure whether that was considered *working near*? *Large* is vague; some respondents categorized a drill press as large, while other respondents did not.) Do you need the help of other persons in handling personal care needs? (Respondent did "get help" but did not "need help"; they were unsure which answer was most appropriate.) Did you have sex with your current partner? (Respondents did not know what the term sex was referring to→Could refer to intercourse or other forms of intimate contact; their answers varied depending upon interpretation.)

Overly Complex	**Question is long or overly complex** • Respondents are unable to remember or synthesize the details in the question because of the length • Overuse of qualifiers, such as "including" and "not counting" • Important ideas are missed because they are buried within the question • Unnecessary grammatical complications (e.g. double-negatives) or awkward structure	Do you have any <u>difficulty</u> hearing, seeing, communicating, walking, climbing stairs, bending, learning or doing any similar activities? (Respondents had trouble remembering examples) What kind of health insurance or health care coverage do you have? Include those that pay for only one type of service (nursing home care, accidents, or dental care). Exclude private plans that only provide extra cash while hospitalized. (Respondents had trouble with use of qualifiers and misreported) In the past 12 months, how many times have you seen or talked on the telephone about your physical, emotional or mental health with a family doctor or general practitioner? (Many did not hear "seen" and responded with only the number of times spoken on the telephone)
Assumption/ Double Barrel	**Question makes inaccurate assumption about respondents' experiences** **Question asks about two different concepts, but assumes a single answer**	Do you think women and children should have universal health care coverage? (Respondents who thought children should, but not necessarily women, were unsure how to answer.) As a result of this injury/poisoning, how much work did you miss? (Respondents who were jobless misreported by saying that they did not miss any work→they had no work to miss)

RETRIEVAL		
Recall/Estimation Difficulty	**Respondent does not know or has difficulty recalling information.** **Response is subject to error because of calculation or estimation difficulty.** The difficulty may be because of memory trouble or because the respondent never knew the information to begin with. • Topic is of low salience to R (especially for attitude questions) • Question asks about information the respondent is unlikely to know • Telescoping: Specific recall problem where events before the reference period are recalled as being within the reference period. • Requires broad reliance on estimation (a wild guess is the best possible result). • Guessing or assuming the answer – ballparking • Trouble making calculations in head	What is the dosage of your asthma medication? (respondents had difficulty recalling dosage details) In your lifetime, with how many people have you had sex? (respondents had difficulty calculating response) Do you think the Lt. Governor of Delaware has been doing a good job? (respondents had difficulty because they did not know information) On what date did you first ride a bicycle? (respondents had difficulty due to recall issues) Asking floor manager about hours that the office staff works. (respondent did not have enough information to answer question)
JUDGEMENT		
Biased/Sensitive	**Response process is influenced by either the sensitive or social context of the question topic.** **The format of the question text or response options produces unintentional influences on the response process.**	Are you in favor of allowing the interests of big business to dominate the Congressional agenda, or are you in favor of sensible legislation to moderate the influence of outside lobbyists? (respondents felt pressured to select second option) With how many men have you had sex? (respondents felt uncomfortable with answering this question) People in the United States are from many countries, tribes and cultural groups. What is YOUR/NAME'S ancestry or tribe? For example, Italian, African American, Dominican, Aleut, Jamaican, Chinese, Pakistani, Salvadoran, Rosebud Sioux, Nigerian, Samoan, Russian, etc. (Rather than viewing the list of ancestries as examples, respondents viewed the list as an exhaustive list of choices)

Questionnaire Effects	**Question problem related to the questionnaire** Question would not necessarily have a problem independently, but there is a problem related to the entire questionnaire. • Illogical juxtaposition of topics/awkward flow • Response of question is impacted by previous questions (i.e. context or order effects) • Inconsistent verbal format across similar or related questions • Questions are perceived to be repetitive • Burdensome length of the questionnaire • Apparent omissions that confuse or irritate respondents.	Do you have asthma? Do you have chronic bronchitis? Do you have coronary heart disease? What was you total household income before taxes in 2002? (The last question in this series appears to be out of context and placed additional burden on respondents) Series of questions about exercise where "no" responses start to create an image of inactivity that makes the respondent uncomfortable; questions that respondents seem reluctant to answer In the 2 years before your first positive HIV test, how many times were you tested for HIV? How often did you get tested for HIV in the 2 years before your first positive HIV test? (these two questions appeared to be repetitive to respondents) Respondent complains about length or does not concentrate to provide accurate information In a questionnaire for truck drivers about toxic exposures, R comments that there were no questions about exhaust from other cars, which is one of the worst problems.
RESPONSE		
Inadequate Response Options	**Response options are incomplete, inadequate or overlapping** • R cannot adequately express their response given the categories provided • Non-mutually exclusive response options; two categories could apply, but only one can be selected • Response options do not fit the question • Response units are inappropriate • Response categories include a problematic term, are ambiguous, or are overly complex • Length of response options places burden on respondent, causing respondents to overlook appropriate responses or to choose the first response that appears to be appropriate without considering subsequent responses	Do you NOW smoke cigarettes every day, some days or not at all? (Many people wanted to report "many days" because "some days" seemed like not enough) What was the reason that you had a mammogram?[Pick one: Regular check up, doctor suggested it, had concern that there was a problem, thought it was time to get one.] (respondents felt the list of reasons were inadequate) Does a long-term physical condition or mental condition or health problem, reduce the amount or the kind of activity you can do at home? Sometimes. Often or Never. (respondents felt response options were inappropriate) How many canisters of asthma medications do you use per month? _____ Number of canisters (Most respondents used no more than one canister per month)

TRANSLATION		
Semantics/Syntax	**Collapsed code. Use this code for any problem that fits the four Codes immediately below.**	
Word selection/connotation	**The word or phrase selected for translation does not accurately or fully convey the English-language intent. This includes translation errors.**	Use the word "LIVIANAS" instead of "ligeras." "Ligeras" implies speed." This "Last" does not mean most recent. It literally means the very last pack purchased. This question uses the word "aumentan" while question 3 uses "incrementan." Which word is preferred?
Grammar	**Translation is grammatically incorrect.** • **Subject/verb do not agree** • **A word is omitted** • **Punctuation problem** • **Incorrect form of a modifier**	Omitted the word "a" in "…Por favor incluya… o para llegar **a** algun…" "Or" has been left out.
Formality	**Formal address when informal should have been used, or vice versa.**	Use "Por lo general,…" I feel it is more familiar speech. "En general,…" is more formal." The words "aùp duïng bôûi" (applied by), [too] sophisticated/educated word for common people.
Awkward construction	**Translation is correct but awkward.** • Sentence construction adheres more to English grammar rules than to those of the target language. • Words or phrases are inserted that make the question sound awkward (e.g., translated item is acceptable in written form but does not flow smoothly when spoken)..	Translated too literally. Translation has too many words, which makes the sentence awkward. The words "aùp duïng bôûi" (applied by), it's a passive voice.
Cannot be translated	**There is no accurate way to convey the English-language concepts in the target language.**	Response category "Algo bien…" is awkward because it is not natural language. Consider using "un poco bien" even though it is not really natural language either. This is a unique name for potato dish (home fries, hash browns); most of the respondents will recognize it in English name rather than to explain

		it in Chinese. Therefore an English name was added next to it.
OTHER		
Evident problem/insufficient information to code	**A problem is indicated in the comments about the item, but there is not enough information to assign a specific code to it.**	Use "son" not "estan." Consider using "…incluya las veces que ya ha mencionado." I would use "Por la mayoria…" instead of "En su mayoria…" "Me siento…" versus "Me hacer sentir…" Add "ser" to the prompt.
Typos	**A typographical or formatting error was discovered in either the source or the target language document.**	Type in English response category 4 "No**t**…"

Coding Guidelines

- Assign one code from each level per problem.
- A questionnaire item can have more than one problem, which means one item can have more than one code per level.
- Be as specific and descriptive as possible when using the detailed Q-Bank codes.
- Use the translation codes sparingly for the cognitive interview documents.
- When coding the "recommendations" section of the cognitive interview documents, do not code any problems mentioned there that have already been coded in the results section. Only assign codes to the recommendations section when an additional and new problem is identified there.

References

Brislin, RW (1970). Back-translation for cross-cultural research. Journal of Cross-Cultural Psychology, 1, 185-216.

Carrasco, L. (2003). The American Community Survey (ACS) en Espanol: Using cognitive interviews to test the functional equivalency of questionnaire translations. (Study Series, Survey Methodology # 2003-17. Washington, DC: Statistical Research Division, U.S. Census Bureau.

Edwards, WS. , Fry, S, Zahnd, E, Lordi, N, Willis, G & Grant, D (May, 2003). Behavior coding across multiple languages: The 2003 California Health Interview Survey as a case study. Paper presented at the American Association for Public Opinion Research. Nashville, TN.

European Social Survey. 2007. Round 3 ESS translation strategies and procedures. Available at http://www.europeansocialsurvey.org/index.php?option=com_content&task=view&id=66&Itemid=112. Accessed June 15, 2007.

Fiscella, Kevin; Franks, Peter; Doescher, Mark P.; and Saver, Barry G. (2002). Disparities in Health Care by Race, Ethnicity, and Language Among the Insured: Findings From a National Sample. Medical Care. 40(1):52-59.

Forsyth, BH, Kudela, MS, Lawrence, D, Levin, K & Willis, GB (in press). Methods for translating an English-language survey questionnaire on tobacco use into Mandarin, Cantonese, Korean, and Vietnamese. Field Methods.

Forsyth, BH, Levin, K, Norberg, A, Thompson, FE & Willis, GB (under review). Using cognitive interviews to test Spanish-language versions of dietary questions.

Harkness, JA (2003) Questionnaire translation. In Harkness, JA, Van de Lijver, FJR, & Mohler, PPH (Eds) Cross-Cultural Survey Methods. Hoboken, NJ: Wiley

Harkness, J.A., Van de Vijver, F.J.R., & Mohler, P. (2003). Cross-cultural survey methods. New York: Wiley.

Hunter, J. & Landreth, A. (2006). Behavior coding analysis report: Evaluating bilingual versions of the non-response follow-up (NRFU) for the 2004 Census test; Study Series Survey Methodology #2006-7; Report prepared for the Statistical Research Division of the U.S. Census Bureau, Washington DC; Issued August 30, 2006.

Kudela, M.S., Forsyth, B.H., Levin, K., Lawrence, D., Willis, G. 2006. Cognitive Interviewing versus Behavior Coding. Paper presented at the American Association for Public Opinion Research, Montreal, Quebec, Canada, May 18-21.

McKay, R. B., M. J. Breslow, R. J. Sangster, S. M. Gabbard, R. W. Reynolds, J. M. Nakamoto, and J. Tarnai. 1996. Translating survey questionnaires: Lessons learned. New Directions for Evaluation 70 (Summer): 93–105.

Miller, K. (2006). Q-BANK: Development of a Tested-question Database. Proceedings of the Section on Government Statistics, American Statistical Association, pp. 1352-1359.

Miller, K., Willis, G., Eason, C., Moses, L., & Canfield, B. (2005). Interpreting the results of cross-cultural cognitive interviews: A mixed-method approach. *ZUMA-Nachrichten Spezial, Issue #10*, pp. 79-92.

Napoles-Springer, AM, Santoyo-Olsson, J, O'Brien, H & Stewart, AL (2006). Using cognitive interviews to develop surveys in diverse populations. Medical Care, 44 (11, suppl 3), S21-S30.

Schoua-Glusberg, A. (2006). Eliciting education level in Spanish interviews. Paper presented to the American Association of Public Opinion Research, Montreal, Canada.

U.S. Census Bureau (2004). Census Bureau Guideline: Language translation of data collection instruments and supporting materials. Washington, DC.

Willis, G., Lawrence, D., Kudela, M., & Levin, K. (2005a). The use of cognitive interviewing to study cultural variation in survey response. Paper presented to the Quest Workshop on Questionnaire Design and Testing.

Willis, G., Lawrence, D., Thompson, F.E., Kudela, M., Levin, K. & Miller, K. (2005b). The use of cognitive interviewing to evaluate translated survey questions: Lessons learned. Proceedings of the Federal Committee on Statistical Methodology Research Conference. Arlington, VA.

Willis, GB, & Zahnd, E. (May, 2006). Response Errors in Cross-Cultural Surveys. Paper presented at the American Association for Public Opinion Research. Montreal, Quebec, Canada, May 18-21.