Annotated Bibliography

Guidelines for Translating Surveys in Cross-Cultural Research

Prepared by the Measurement and Methods Core
of the Center for Aging in Diverse Communities,
University of California San Francisco

Introduction

As the population become more diverse, it is imperative to conduct health research within non-English speaking populations. Only recently have health researchers begun to identify best practices for the translation and assessment of translations of survey instruments into other languages. Current standards for translation procedures are lacking and few researchers report their methods of translation. Often, time and money are not dedicated to the proper translation and adaptation of measures, and thus the cultural and conceptual equivalence often suffers.

A well-translated survey instrument should have semantic equivalence across languages, conceptual equivalence across cultures, and normative equivalence to the source survey. Semantic equivalence refers to the words and sentence structure in the translated text expressing the same meaning as the source language. Conceptual equivalence is when the concept being measured is the same across groups, although wording to describe it may be different. Normative equivalence describes the ability of the translated text to address social norms that may differ across cultures. For example, some cultures are less willing to share personal information or discuss certain topics than other cultures. If possible, both surveys should be developed simultaneously, preventing the survey from being based too deeply within one culture and language. Furthermore, some researchers have begun to consider whether the same questions should be asked of all populations, or whether cultural considerations may require slightly different questionnaires in several cases (issues specific to religion, health beliefs, etc).

Below we provide a list of journal articles, books, and book chapters that describe recommended methods of translation of survey instruments into multiple languages in cross-cultural research.

KEY ARTICLES AND REPORTS THAT DISCUSS METHODS FOR TRANSLATING SURVEYS INTO MULTIPLE LANGUAGES

Beaton DE, Bombardier C, Guillemin F, Ferraz MB. Guidelines for the process of cross-cultural adaptation of self-report measures. *SPINE* 2000; 25 (24):3186-3191.

This article provides a concise guide to adapting self-report measures for cross-cultural use. The authors suggest a five stage process of translation, synthesis, back translation, expert committee review, and pretesting. According to this source, translation should involve at least 2 independent forward translations by bilingual translators which can be compared to identify discrepancies indicative of ambiguous wording within the original survey or other problems and revised accordingly. During synthesis, a third bilingual person mediates a discussion between the two translators to develop one version of the survey. Written documentation of the process is encouraged. Another person blind to the original survey then back translates the new survey into the source language and compares it to the original document to check the validity of the translation. An expert committee, comprised of the translators and health and language professionals, meets with the purpose of consolidating the different versions of the survey to produce a final form and ensure equivalence between the source and new versions. The translated survey should then be pretested in a sample of 30-40 persons from the target population using standard cognitive interviewing techniques. The authors consider testing of the final translation in a larger sample as a distinct step from translation and therefore do not cover it here.

Behling O, Law KS. *Translating Questionnaires and Other Research Instruments: Problems and Solutions*. Thousand Oaks, CA: Sage Publications Inc. 2000: pp 1-63.

This booklet provides an overview of key issues involved in the translation of questionnaires including achieving semantic equivalence across languages, conceptual equivalence across cultures, and normative equivalence across societies. The authors explore these three levels of equivalence and the problems one may have at each level across different types of questions asked (demographic, behavioral reports, knowledge, etc). For example, it is relatively easy to achieve semantic and conceptual equivalence of demographic questions across languages, since the words and ideas are more general and commonly used. However, it is harder to achieve normative equivalence, since cultures differ on how willing they are to share personal information. On the other hand, it is much more difficult to achieve all types of equivalence when translating and asking questions about attitudes and opinions since the ideas are more abstract, the concept may not be relevant throughout the world, and some cultures may resist discussing certain beliefs with strangers. The authors review and rate 5 methods often used to establish semantic equivalence when translating a survey from an existing survey including direct translation, back translation, and random probes. Practical advice is also given for achieving semantic equivalence when creating a new survey including writing with translation in mind, decentering, and using multicultural teams. Empirical tests that can be used to test conceptual equivalence of survey items (factor analysis, item response theory) are discussed. Normative problems that can arise in cross cultural research include social norms about openness with strangers, political opinions, tendency to conform or assert oneself, and more. The authors provide several ideas for addressing these issues: develop close relationships with respondents or use individuals who are trusted within the sample to recruit or interview for the survey; use multicultural teams when translating the survey; and pilot test the survey.

Bullinger M, Alonso J, Apolone G, Leplege A, Sullivan M, Wood-Dauphinee S, Gandek B, Wagner A, Aaronson N, Bech P, Fukuhara S, Kaasa S, Ware JE Jr., for the IQOLA Project Group. Translating health status questionnaires and evaluating their quality: The IQOLA Project approach. *J Clin Epidemiol* 1998;51(11):913-923.

This paper describes the methods adopted by the International Quality of Life Assessment (IQOLA) project to translate the SF-36 Health Survey. These methods continue to be used with translation of the SF-36 into previously unavailable languages. The IQOLA developed a three-stage process to produce cross-culturally comparable translations of the SF-36: 1) rigorous translation and evaluation procedures to ensure conceptual equivalence and respondent acceptance; 2) formal psychometric tests of the assumptions underlying item scoring and construction of multi-item scales; and 3) examination of the validity of scales and the accumulation of normative data. This article focuses on the first step, translation. Methods used included forward and backward translation by at least two translators, translator ratings of difficulty of translating an item and quality of translation, pilot testing, and cross-cultural comparisons of translations. Techniques that contributed to improvements of the translations included reworking translations with low quality ratings, comparing backward translations with the original SF-36 questionnaire, and cross-cultural discussions about the translations of items and response choices.

Forsyth BH, Kudela MS, Levin K, Lawrence D, and Willis GB. Improving questionnaire translations and translation processes. Paper presented at Q2006, European Conference on Quality in Survey Statistics, April 25, 2006. Cardiff, Wales, UK.

This paper explores procedures for developing and evaluating questionnaire translations for surveys administered in multiple languages. The authors focus on a case study in which they translated an English-language questionnaire on tobacco use into Mandarin, Cantonese, Korean and Vietnamese. A team of three translators each translated the survey from English into one of the target languages, and kept

detailed records of the specific translation challenges they encountered and the decisions they made to deal with the challenge. Four survey language consultants (SLC) were hired to review the new translations and coordinate pretesting activities. A formal process was established by which SLCs could review the survey, identify problematic areas, document their findings and suggest a revision. These written documents were used in the final adjudication phase. The survey translations were pretested using cognitive interviews, and the results were used to make final changes to the surveys. The authors detail the five step translation, evaluation and review process they used and the lessons learned at each step. Some of the important lessons learned include engaging survey reviewers early during translation to reduce the need for large-scale revisions later on, and provide translators with unambiguous instructions, including the reasons for and structure of the survey interview.

Hagell P, Hedin P, Meads DM, Nyberg L, McKenna SP. Effects of method of translation of patient-reported health outcome questionnaires: A randomized study of the translation of the Rheumatoid Arthritis Quality of Life (RAQoL) Instrument for Sweden. *Value Health* 2010 Jan 8 [Epub ahead of print].

This paper is unique in that it consists of a randomized trial comparing the quality of an instrument that was translated using two independent translation methods: forward-backward translation (FB) and dual-panel methods (DP). In the forward translation version of the instrument, two forward translations were combined into one Swedish version by the authors taking into account conceptual considerations. This version was back-translated into English by a third translator. The Swedish version was then assessed by 10 people with rheumatoid arthritis (RA). In the dual-panel method, a panel of six bilingual Swedes working with one of the instrument developers produced a draft Swedish version, which was then reviewed and revised by a 2nd panel consisting of six monolingual Swedes who did not have RA. This was followed by a face-to-face field test with 15 people with RA, but no changes were made since interviewees reported no problems with the questionnaire. 200 RA patients were then randomized to take the FB or DP version. There were more missing items with the FB than DP version (6.9% vs. 5.6%; $p<.0001$); reliability was .92 for both versions. Qualitative ratings were completed by 11 lay people, 23 bilingual Swedes, and 50 people with RA. Lay people and patients preferred the DP over the FB item versions ($p<.0001$). Construct validity was similar for both versions. Differential item functioning by version was found for five items, but did not affect estimated person measures. Findings suggest that the two versions demonstrated similar psychometric properties, but the DP approach showed advantages over the FB translation from the patients' perspective. This paper supports the need for systematically testing various survey translation methods.

Harkness J, Pennell BE, Schoua-Glusberg A. Survey Questionnaire Translation and Assessment. In: Presser S, Rothgeb J, Couper M, et al. *Methods for Testing and Evaluating Survey Questionnaires.* Hoboken, N.J.: John Wiley & Sons, Inc. 2004:453-473.

This chapter identifies several key difficulties when translating existing surveys including maintaining the intended meaning of the questions and matching the semantic content and structure across languages in both questions and answer scales. Existing survey questions may be slightly ambiguous as to their intended meaning, forcing translators to either leave their translation ambiguous or decide on a single interpretation and translate the survey accordingly. This allows for the possibility of different meanings of questions across languages. The authors provide examples of common problems that arise when a questionnaire is translated too closely, meaning the translation focuses on the words and not the meaning of the questions. Possible problems include creating a different question than the original, creating an unnecessarily complicated or awkward text, and the unidiomatic or improper use of the target language. The authors also lay out issues regarding the translation of answer scales. For example, in some languages the difference between 'disagree' and 'not agree' does not exist, and therefore response

options must be altered from the source language. The authors detail the benefits of using a team approach to translation and review, and outline several qualitative (cognitive interviews, interviewer and respondent debriefing, back translation) and quantitative (statistical tests) approaches that can be used in the review process. The chapter concludes with an emphasis on the type of documentation necessary for a successful translation including background documentation that should be provided to translators, record keeping of changes made to the survey, and documentation of final versions.

Harkness J. Questionnaire Translation. In: Harkness J, Van de Vijer F, and Moher P. *Cross-Cultural Survey Methods.* Hoboken N.J., John Wiley & Sons, Inc. 2003: 35-56.

This chapter outlines the practical implementation and assessment of questionnaire translation, emphasizing important procedures for translating questionnaires and the staff skill set necessary to complete the task. The authors argue that 3 sets of people are necessary to translate a survey: translators, translation reviewers and translation adjudicators. Each group will have varying degrees of knowledge and training with the target language, translation skills, principles of research, and the study design depending on their role. The authors outline the TRAPD model (Translation, Review, Adjudication, Pretesting and Documentation) and details how a committee approach can be used within the model. Both parallel translation (several translators do independent translation) and split translation (multiple translators translate different sections of survey) are acceptable methods. A committee then reviews the entire survey, discusses the translation, and decides upon a version. Adjudication (deciding on the final version) can be done at the same time or by a separate committee in a second round of review. If the survey is complicated, often expert consultants can be brought in at this step to assist with finalizing the survey. Extensive assessment and pretesting is crucial to produce a quality translation of a survey. Pretesting techniques that can be used include review of the survey by focus groups, cognitive pretests, back translation and soliciting feedback from field staff, monolingual and bilingual respondents. Documenting the translation process assists the reviewers and adjudicators in developing the final version of the survey, and keeps a record of any adaptations that were made between the different languages.

The chapter also discusses aspects of linguistics directly relevant to questionnaire translation including issues of gender and sentence structure, close or literal translations, and the translation of response scales. This book also includes chapters that deal with other issues relevant to developing, conducting and analyzing survey research in cross cultural populations including sampling, bias and equivalence, non-response, social desirability, and data collection methods.

Hunt SM, Bhopal R. Self report in clinical and epidemiological studies with non-English speakers: the challenge of language and culture. *J. Epidmiol Community Health* 2004; 58: 618-622.

In this article, Hunt and Bhopal argue that when data collection instruments designed for English-speakers are translated into other target languages, there are often measurement errors due to poor translation procedures, inappropriate content, insensitivity of items, and a lack of knowledge of the cultural norms by researchers. Traditional translation methods involve a bilingual professional translating an English document into the target language, focused on achieving linguistic equivalence. Pretesting such surveys has shown that bilingual professionals are not representative of the sample population, and often produce translations that are too formal. The author suggests several translation and adaptation procedures to overcome these shortcomings including consulting and field testing measures within a monolingual sample of the target population and testing for face, content and construct validity in each language. Even extensive testing cannot always create perfectly equivalent items in several languages due to the fact that often there are no equivalent terms for a given concept. Culture must be considered when developing the survey. For example, western ideas of risk, health and need may not be as dominant in other cultures that have alternative views. The authors suggest that possible approaches to improving

cross-cultural surveys include developing both emic and etic questions around a given topic and developing equivalent concepts instead of equivalent items.

Keller SD, Ware JE, Gandek B, Aaronson NK, Alonso J, et al. Testing the Equivalence of Translations of Widely Used Response Choice Labels: Results from the IQOLA Project. *J Clin Epidemiol* 1998; 51 (11):933-944.

This article describes a study conducted by the International Quality of Life Assessment (IQOLA) Project which tested the relative magnitude scaling of response choice labels of the SF-36 across languages and countries. The project evaluated whether the 1) ordinal values of response options (*does very good fall between fair and excellent?*), 2) interval difference between response options, and 3) the translation of response labels were equivalent across countries. Results indicate that the ordinal value of response options assigned by respondents mostly matched that of the current SF-36. Labels such as 'a good bit of the time, some of the time, and most of the time' and response options involving the term 'moderate' are examples of where different ordinal values were sometimes assigned to response options. Overall, the numerical scores assigned to the SF-36 response options were replicated by the current study. In some cases the distance between terms such as 'very good' and 'good' were scored as closer together than the current assigned values. Generalizability of response labels and their translations across countries was supported by the results.

Ponce NA, Lavarreda SA, Yen W, Brown ER, DiSogra C, Satter DE. The California Health Interview Survey 2001: translation of a major survey for California's multiethnic population. *Public Health Reports* 2004;119:388-395.

This article describes the process used to translate a population-based telephone health survey into Spanish, Chinese, Vietnamese, Korean, and Khmer. Cultural adaptation was conducted by a statewide panel of 12 bilingual-bicultural reviewers with survey expertise focused on the targeted ethnic groups. These cultural experts independently rated each question in English on a 4-point scale from 1=problematic item to 4=exemplary item. Results were then discussed by the survey team and reviewed in focus groups in English with African Americans, and American Indians/Alaska Natives; items were culturally adapted based on the results. For translation, CHIS used translation by committee or "multiple forward translations (MFT)," as they prefer to call it. MFT consists of translators creating two or more forward translations, which are then reconciled by another independent translator. They also used an outside referee to judge the quality of each of these translations or refereed multiple forward translation (RMFT). The proportion of respondents interviewed using a translation in each of the targeted groups ranged from 34% to 50% supporting that translation of surveys is essential for adequate representation of these ethnically diverse groups in population-based surveys in California.

Smith, T.W. Developing and Evaluating Cross-National Survey Instruments. In: Presser S, Rothgeb J, Couper M, et al. *Methods for Testing and Evaluating Survey Questionnaires*. Hoboken, N.J.: John Wiley & Sons, Inc. 2004: 431-452.

This chapter describes common challenges that arise when creating and conducting cross-national surveys including maintaining equivalence of wording and meaning of questions, maintaining equivalence of answer response scales, and response effects. The goal of developing questions that function equivalently across languages is hindered by words that have no equivalent translation, or words that have linguistic equivalence but represent slightly different concepts (i.e. the concept of equality/égalité is different in the U.S., English-speaking Canada and French-speaking Canada). The problems related to linguistic and conceptual equivalence can be addressed by using multiple indicators for each construct. The author argues that by using at least 3 linguistically distinct measures (use multiple items, each item using different terms for the same concept) one can discern if the items and construct works equivalently across the different languages. Developing answer response scales that are equivalent

is the second step in the process. Many researcher support the use of numerical scales to increase equivalency across groups, however, this can be problematic because numerical scales are often complex, people tend to use a small percentage of the numbers on the scale or avoid the extreme values, numbers can be considered lucky or unlucky, and numerical scales must still be described and anchored by a verbal term that must be equivalent across languages. Other methods that can help create equivalent answer scales are asking respondents to calibrate the scale by rating each response option term on a numerical scale, or have respondents answer the same question several times with different sets of response options to directly compare response options. Cross-national data can also be compromised due to response effects that differ across groups such as the tendency to select socially desirable response options, select or avoid extreme response options, and the tendency to select neutral or middle options. Mode of survey administration and question order will also impact survey data. To deal with these types of issues, survey development should occur through a collaborative multinational approach which takes translation into account during the development of the survey, not as an after thought. The author provides practical advice for designing questions which use simple language to allow for easier translation. Other issues to consider when conducting cross-national research such as including emic and etic questions are explored.

Solano-Flores W, Hurtado M, et al. CAHPS Guidelines for Assessing and Selecting Translators and Reviewers. CAHPS II Cultural Comparability Team, Jan 2005.

This document provides guidelines for the assessment and selection of translators and translation reviewers used for the CAHPS survey. The authors address three major topics: the roles of the translator and the translation reviewer; the process of selecting translators and translation reviewers (or translation firms); and the qualifications that each should have. The translator's role is to produce a translated text that is accurate, grammatically correct, sensitive to regional variations and written at an appropriate reading level. Translator reviewers (often a committee of reviewers) check the work of the translators to ensure that the text is accurate, written at an appropriate level and that all technical terminology is correct and understood by the majority of people. Translators and reviewers should be native speakers of the target language, proficient in the reading the source language, experienced in translating documents and have experience within the health services field.

US Census Bureau, Census Bureau Guideline: Language Translation of Data Collection Instruments and Supporting Materials. Census Bureau Website. Accessed March 20, 2007. http://www.census.gov/cac/www/007585.html

On this website the Census Bureau outlines the methodology they use to translate survey instruments and provides several attachments and supporting documents including an overview of the methods and current state of knowledge, criteria for achieving good translations, and a sample translation validation form. The website details the 5 step protocol the Census Bureau follows and recommends: Prepare, Translate, Pretest, Revise and Document. The Census Bureau does not recommend solo or direct translation with back translation, but instead strongly promotes a process of translation and review by a team of translators, reviewers and adjudicators. At a minimum the team should include 2 translators to perform the translation, an expert in the subject matter, a person knowledgeable in survey design and an adjudicator. As preparation, translators should be supplied with a summary of the scope of the project, explanation of the target audience and survey mode, survey documentation that provides definitions of terms or concepts, and access to people who can assist them with questions about the subject matter or questionnaire design. Pretesting is a necessary step that identifies problems in the translated text or helps identify other concepts that may be relevant within the target population. Documentation of the translation process at each step makes it possible to track the different survey versions or demonstrate that the survey functions well in the pretests. Many of the guidelines the Census Bureau presents stem from a two-day

expert panel meeting, which was designed, sponsored, and hosted by the Census Bureau in November 2001.

The documents provided as attachments on the website give researchers more practical guidance in developing surveys in other languages including things to consider (audience, geographic location, social and cultural factors) that will impact the nuances of the translation. The supporting document: **Translation of Surveys: An Overview of Methods and Practices and the Current State of Knowledge** provides a brief review of the current state of knowledge of developing questionnaires in multiple languages and presents several of the most commonly used approaches including direct translation, back-translation and committee approach. These resources are also compiled in the following report by the Statistical Research Division: Pan Y, De la Puente M. **Census Bureau Guideline for the Translation of Data Collection Instruments and Supporting Materials: Documentation on How the Guideline Was Developed.** 2005 (August 24). Washington, D.C.: U.S. Census Bureau.

Willgerodt MA, Kataoka-Yahiro M, Kim E, and Ceria C. Issues of Instrument Translation in Research on Asian Immigrant Populations. *J. Prof Nurs* 2005; 21(4):231-9

The article describes the seven steps of the Brislin translation method with decentering (described in Werner and Campbell 1970) and documents the authors experience implementing the procedures in studies of two Asian immigrant populations (Filipino Americans and Korean Americans). In the Brislin method the questionnaire is translated and back translated independently by two translators, reviewed by a team and pretested in a sample of the target population. Following the pretest, the survey is administered to a group of bilingual subjects; some receive the English version, some receive the target language version, and some receive both. The means, standard deviations and correlation coefficients for all versions are compared.

Using the Brislin method, the authors describe several issues encountered in trying to achieve semantic and content equivalence in two separate samples. In the Filipino study, translators of the Caregiver Reaction Assessment (CRA) had difficulty developing equivalent terms for words such as 'resent', 'financial strain', 'want' and 'enough'. Cultural differences in the concepts of family and care giving were made apparent during the pretest, and other concepts regarding the family that are important in Filipino culture were not included in the original English measures, bringing the measures applicability into question. In the Korean sample, a literal translation of the Parenting Practices Interview (PPI) created problems in sentence structure in the Korean version, and a more liberal translation was necessary. Problems also arose translating the concept of ignoring bad behavior as a disciplinary strategy for children, and the review team went through many translation and back-translation cycles before deciding on a term. The authors provide many practical recommendations for translating instruments for use within Asian immigrant populations. Major conclusions from these studies include the necessity of a skilled translation and review team with bilingual experts familiar with the study content and everyday language and culture of the target population, the need to evaluate the original instrument for cultural relevance (does it include all aspects of the construct it is measuring relevant to the population), and the need to pilot test the translated measure to identify problems and develop semantic and content equivalence.

Willis G, Lawrence D, Thompson F, Kudela M, Levin K, and Miller K. The Use of Cognitive Interviewing to Evaluate Translated Survey Questions: Lessons Learned. Paper presented at 2005 Conference of the Federal Committee on Statistical Methodology. November 14, 2005. Arlington, VA.

The authors propose cognitive pretesting as a necessary step in the translation process of multi language survey instruments. This paper uses three case studies to demonstrate how cognitive pretesting can assist in developing improved multi-language survey instruments by identifying translation errors and

culture specific and general problems within the instrument. In the first case study, cognitive pretesting provided information that would not have been captured by the standard survey administration. For example, many respondents who reported that they would not walk to the store in a rainstorm to buy cigarettes indicated after further probing during the pretest that this would not be necessary since they always kept enough cigarettes on hand to ensure they never ran out. The second case study demonstrated that following the probe script too closely or robotically during the pretest is problematic and asking general comprehension probes ("What does this questions mean?") made subjects feel as though they were being tested. Specifically targeted probes ("Did you include X when you were thinking about the fruits that you ate?") were most useful.