CJNR 2003, Vol. 35 Nº 2, 127–142

Designer's Corner

Translating and Adapting Measurement Instruments for Cross-Linguistic and Cross-Cultural Research: A Guide for Practitioners

Elizabeth A. Kristjansson, Alain Desrochers, and Bruno Zumbo

The psychometric instruments used in cross-linguistic or cross-cultural research are typically developed in one language and then translated into another. The authors address methodological problems that arise in the translation process and that compromise data interpretation, using concrete examples to illustrate these problems. They point out and describe the relevance of lexical semantics in item translation. The authors make recommendations for avoiding common pitfalls in the use of translated measurement instruments or in the translation or adaptation of such instruments. This paper is intended for researchers who are planning to develop or use translated instruments.

Measurement is integral to nursing practice. It is used to facilitate the diagnosis of physical and psychological health problems and in the assessment of pain as well as physical and cognitive functioning. It can also serve a useful purpose in the assessment of self-reported health status, knowledge or attitudes about illness, and health services or policies. Measurement instruments may take a variety of forms, including questionnaires, tests, rating scales, and self-reports. The choice depends largely on the purpose of the research that is undertaken. Psychometric measures typically hinge upon a focal theme or concept. For instance, a performance test may be designed to assess a cognitive ability such as attention or memory. An attitudinal scale may quantify individuals' disposition towards smoking or their impression of services dispensed by a healthcare institution. A questionnaire may serve to assess older persons' quality of life. All themes or concepts generally involve various facets, which are measured using different questions or items. The potential range of applications of such measures is open-ended.

Researchers take a number of precautions to ensure that the conclusions they draw from their data are sound. For instance, instruments must include a sufficient number of items to represent the relevant facets of the concept being measured. The responses to the items that are conceptually related are expected to be correlated and form a consistent body of data. Over multiple measurement sessions, these responses must reflect accurately the stability of or a change in what is being measured (e.g., an ability, an attitude). Psychometric measures should also be correlated with other, conceptually related, measures, and be uncorrelated with conceptually unrelated measures. More importantly, they are expected to provide a fair and unbiased representation of the underlying concept. Meeting these conditions is essential to establish the reliability and validity of a measurement instrument.

Item and test bias is a potential threat to the validity of the inferences made from psychometric measures. The term "bias" is used here to refer to a systematic error in the measures taken from a group of individuals. When an item or an instrument is biased against one socio-economic, linguistic, or cultural group, the measures derived from it do not accurately reflect a person's true abilities or characteristics (Camilli & Shepard, 1994). Such a measurement error may be due to some characteristic of an item or situation that is not relevant to the purpose of the instrument. It may be introduced, perhaps unknowingly, when the instrument is being constructed, translated into another language, or adapted to a culture or context for which it was not designed.

Measurement problems that occur in the translation and adaptation process are particularly relevant to multilingual and multicultural countries like Canada. Some measures are routinely used to collect data on the health status, attitudes, level of satisfaction, and preferences of English-speaking and French-speaking Canadians. Whenever differences are detected in the data derived from an instrument and its translation into another language, one hopes they reflect differences between the groups of respondents rather than between the two versions of the instrument. Although translating a measurement instrument into another language may seem a straightforward exercise, it is in fact quite difficult and has historically been fraught with problems (Banville, Desrosiers, & Genet-Volet, 2000). For instance, translating a word by its dictionary equivalent does not necessarily ensure conceptual equivalence (Hambleton, 2002). A word and its dictionary equivalent may differ in their range of meaningful associations, number of possible meanings, emotional valence, familiarity, and so on. These problems make it difficult to ascertain whether observed differences among language groups are real or due to bias (Sirecci, 1997). In turn, such ambiguity makes the interpretation of the results a perilous exercise. Careful attention to issues of linguistic and

cultural equivalence during the development, evaluation, and use of translated or adapted instruments will help to ensure that conclusions drawn from the data derived from the instruments are valid and fair for all linguistic and cultural groups.

This paper has two main objectives: to highlight difficulties in translation that may lead to errors in measurement at the item or scale level, and to provide recommendations on the translation or adaptation of measurement instruments. The considerations presented in this paper may be of interest or value to researchers who are (a) using an existing translation of a measurement instrument, (b) developing their own translation, or (c) developing a new instrument in the target language.[1] If you are planning to use an existing translated measurement instrument, this paper may help you to evaluate it critically and to determine whether it will provide valid, reliable, and unbiased measures. If you are translating or adapting an instrument, it may serve to make you aware of the appropriate steps to take. This paper also contributes to the measurement literature by highlighting the various and subtle ways in which translation can change the meaning of measurement items, even in the best of instruments.

Problems and Pitfalls in Translation

A number of hurdles may have to be overcome in the development of a meaningful and equivalent cross-linguistic or cross-cultural measure (see Behling & Law, 2000; van de Vijver & Poortinga, 1991). Problems leading to bias can be grouped into three types: lack of conceptual equivalence, differences in cultural norms, and lack of semantic equivalence (Behling & Law). We shall give a brief overview of each type, paying special attention to problems of semantic equivalence, as these are quite common and are often easy to identify and resolve.

Lack of Conceptual Equivalence

Most psychometric instruments are intended to assess one or more abilities (e.g., cognitive skills), self-perception (e.g., health-related conditions, depression), traits (e.g., personality), or attitudes (e.g., towards health-threatening behaviours). However, psychological concepts are not necessarily universal (Behling & Law, 2000; Hui & Triandis, 1985). A concept

[1] Researchers may have the choice of using existing scales or developing their own scales. When suitable scales are available and well constructed, it is advisable to use them, as they have already been tested, and developing new scales involves an inordinate amount of work. This advice applies equally to translated measurement instruments. Translation and adaptation is an exacting, time-consuming, and expensive process.

that is meaningful in one culture may not be so in another, and even if it is meaningful, it may be more important or salient in one culture than in another. More commonly, the behavioural manifestations of some concepts will differ among cultures. For instance, culture shapes the way we conceive of health and illness and influences the relative value or importance of symptoms. Such differences have implications for the operational definition of a psychological concept and for the development of a psychometric instrument.

Let us consider some concrete examples. Cognitive functioning is operationalized differently in different cultures; its manifestations are largely influenced by the demands of the social environment in which one lives. In North America, for example, numeracy and literacy are practically indispensable skills in most daily activities, from balancing a cheque-book to reading a recipe or the newspaper. In addition, a great deal of emphasis is placed on efficiency, productivity, and speed of response (Hambleton, 2002). Tests of cognitive ability often reflect these skills and values. However, such tests would not provide meaningful measures for people whose lifestyles and experiences differ dramatically from those of North Americans. Teng (1996) aptly illustrates this point with the example of an illiterate grandmother in a developing country who has little or no experience with testing. This woman would likely do poorly on a North American cognitive test, not because she suffers from dementia but because the skills that are emphasized in these tests are not relevant to her life experience, which comprises cooking, gardening, taking care of animals, and raising children. A more appropriate test of her cognitive abilities would emphasize the skills that she needs in her everyday life.

Depression is also expressed quite differently in different cultures; depressive symptoms that are important in one culture may not be important in another (Edwards, 1995; Guarnaccia, Angel, & Worobey, 1989). For example, Bhtanagar and Frank (1997) observed that depressed elderly South Asian immigrants did not demonstrate guilt feelings or suicidal ideation, symptoms that are commonly shown by people who are depressed in North America. The authors attributed the absence of these symptoms to the fact that such feelings are perceived as socially disgraceful in South Asian societies. In another study, Marsella and White found that people in nine non-Western countries rarely reported classic symptoms of depression such as depressed mood, loss of interest, and sleeplessness (as cited in Curyto et al., 1998). The implication of these results is that the items on depression measures should be relevant to the way in which depression is manifested in that culture. We now turn to a related issue, namely how differences in cultural norms influence responses to an assessment.

Differences in Cultural Norms

Societal norms profoundly influence our attitudes and behaviour (Behling & Law, 2000). This influence can be shown in various ways. Let us briefly consider three examples.

Societies differ in their openness and willingness to discuss certain topics. In extreme cases, respondents to a questionnaire might refrain from answering some questions or fail to report instances of particular symptoms or behaviours. Such reluctance to divulge personal information may make the early detection of distress (e.g., resulting from a physical or mental illness) nearly impossible.

Individuals raised in different social or cultural environments may also differ in their inclination to provide socially desirable responses to please an interviewer. In extreme cases, individuals might consistently portray themselves in a way that makes them look good to others, even if this façade is dissonant with the way they are or feel. The consequence for the measures taken from such individuals is that the data will reflect their ideal self rather then their true self. It is precisely this source of bias that led researchers to develop social desirability scales and explore ways to adjust individual data for social desirability (Crowne & Marlow, 1960). Whether or not such social desirability scales apply equally well across cultures is still an open question.

Different cultures may enforce different norms for responding to particular situations (e.g., the death of a spouse) or internal states (e.g., grief or distress). In some cultures, people may feel free or even encouraged to display their pain or distress openly, while in others such behaviour may be perceived as irresponsible or as a form of weakness and be regarded unfavourably. Such variation makes it particularly difficult, in health research, to establish the symptomatology and, more importantly, to assess the true intensity of symptoms.

Behling and Law (2000) address the issue of differences in cultural norms. The solutions they propose are similar to those discussed in the context of conceptual equivalence. Most of them involve learning about the cultures of interest and exercising sensitivity and care in the development of items (e.g., in choice of words, formulation of questions).

Lack of Semantic Equivalence

In psychometric instruments, meanings are typically conveyed through words, phrases, or sentences. The issue of semantic equivalence thus relates to the mapping of meanings across languages: How can an idea expressed in one language be accurately conveyed in another? Languages may differ in subtle ways in their expressive resources. Words or phrases that are meaningful in one language may have no exact counterpart in

another (Bracken, 1990; Retief, 1988). The grammatical strategies used for expressive purposes in one language may have no equivalent in another (Retief). Another hurdle in translation relates to differences in experience and learning, which in turn lead to differential understanding and interpretation of words and other stimulus materials (Retief; van de Vijver & Poortinga, 1991).

Let us consider each of these problems by examining a few examples from a variety of tests and measurement instruments. At the outset, we would like to underscore the fact that the problematic items we cite are drawn from very carefully developed and validated measurement instruments, and a problem with one or two items does not mean that the tests are invalid. Moreover, these problems with item translation would not have been identified but for the vigilance of the developers and other researchers in checking measurement equivalence. These examples, rather than indicating problems with particular tests or measurement instruments, highlight the need to carefully consider the meaning of translated items for the respondents they are intended for; they also highlight the need for thorough and systematic study of item equivalence.

Problems of lexical mapping occur whenever the meaning of a word or an idiomatic expression does not map exactly that of its dictionary equivalent. Strictly speaking, meanings can rarely be conveyed with precision in translation. However, acceptable approximation can ordinarily be achieved. Lack of exact mapping is easy to demonstrate. For example, consider the relationship between the English word *ball* and the French word *balle*. At first glance, these words may be taken as equivalent, but they are not. *Ball* designates a larger set of referents than *balle*, which represents only spherical objects that can be held in one hand (e.g., a baseball). In French, a large ball (e.g., a basketball or football) is a *ballon*. Now consider a test item in which the respondent must match a short sentence to a drawing of a scene such as a boy holding a basketball. If the English version used the sentence "A boy is holding a ball," respondents would likely give a *true* response. However, if the French version used the sentence "Un garcon tient une balle," respondents would likely give a *false* response.

The preceding case may be so obvious as to be dismissed as a poor example. However, consider this item from the Mini-mental State examination (Folstein, Folstein, & McHugh, 1975) and the Modified Mini-mental State examination (3MS; Teng & Chui, 1987). In an item designed to assess attention, the respondent is asked to repeat the phrase "no ifs, ands, or buts." This is a familiar idiomatic expression to most English-speaking individuals (as in "Do it right now, no ifs, ands, or buts"). However, it presents problems in test adaptation because it has no counterpart in other languages (Teng, 1996). Realizing that direct trans-

lation of this item would not work, the experts who adapted the 3MS for Spanish-speaking North Americans substituted "si no sube, baja" ("if it doesn't go up, it goes down") (Marshall, Mungas, Weldon, Reed, & Haan, 1997). However, in a later item analysis, Marshall and her colleagues found that this item functioned differently for English- and Spanish-speaking test takers and recommended that another expression be used. This highlights the difficulties inherent in searching for conceptually equivalent idiomatic expressions.

The implementation of semantic equivalence is required in most cross-language translations, but there are exceptions. For instance, on the Boston Diagnostic Aphasia Exam (BDAE), respondents are instructed to copy the sentence "The quick brown fox jumps over the lazy dog." This test is designed to assess one's ability to reproduce in writing the 26 letters of the alphabet. Its focus is therefore on the visual form of the item rather than its meaning. In a French translation of this item, the respondent received the sentence "Le petit renard brun s'échappe du chien paresseux." Although this item preserves the meaning of the original, it misses the purpose of the test entirely because it uses only 15 of the 26 letters of the alphabet (Garcia & Desrochers, 1997). This example directs us to stress the distinction between translating and adapting a psychometric instrument for another linguistic or cultural group. The ultimate goal of adapting a measure into another language is to preserve the meaning of the theoretical concept. This goal can often be achieved by translating an item from the source to the target language. In particular circumstances, however, researchers may have to generate a new item, as would be recommended in the "quick brown fox" example.

Another example that highlights difficulties in lexical mapping comes from a study comparing the item equivalence of the American version of the SF-36 health questionnaire (Ware & Sherbourne, 1992) and its Danish translation (Björner, Kreiner, Ware, Damsgaard, & Bech, 1998). In one item from the Physical Functioning subscale, respondents are asked how much their health limited them in walking a *mile*. The Danish translation asked how much their health limited them in walking a *kilometre* (a shorter distance). Although a kilometre is more meaningful to Danes than a mile, this translation resulted in a different benchmark for the question. The Danes reported fewer problems with this task than the Americans who had the same level of overall health (Björner et al.), but they were compared on different criteria. This difference compromised the interpretation of the observed difference between the two language groups on the question (Björner et al.). This is an example of the difficulties faced in trying to develop conceptually equivalent items while maintaining cultural relevance.

Problems of grammatical or syntactic equivalence. Languages differ in the way in which sentences are constructed. The typical word order of an item in one language may not be appropriate for its translation into another language. Consider the following similarity item taken from the Modified Mini-mental State exam (3MS; Teng & Chui, 1987), which was used in English and French in the Canadian Study of Health and Aging (McDowell, 1994). Respondents were asked to describe similarities between pairs of concepts. In the English version, they received "In what way are laughing and crying alike?" In the French translation, they received "En quoi se ressemblent rire et pleurer?" Although the translation satisfies the principle of semantic equivalence, it features the less common verb-subject order rather than the canonical form (i.e., En quoi "rire" et "pleurer" se ressemblent-ils?). Preliminary item analyses showed that the item was more difficult for French-speaking than English-speaking respondents with similar cognitive abilities. This example reminds us that the way in which ideas are expressed in translation does matter, as it can influence performance.

Experiential equivalence. Most psychometric instruments rely heavily on the use of language. Since the interpretation of language usually involves general knowledge, one must also consider experiential equivalence in the translation or adaptation of an instrument.

Ellis (1989) reports a relevant example of differential knowledge in the cross-linguistic equivalence of the Career Ability Placement Survey (CAPS). This questionnaire was developed in English and subsequently translated into German. On one item of the Verbal Reasoning subtest, respondents were given the information that "the dogs in the park are all retrievers" and that "Cindy owns a poodle" (Ellis, p. 921) and were then asked whether the statement "All of Cindy's dogs are in the park" was *true, false,* or *uncertain.* Most North Americans would answer *false,* because one of Cindy's dogs is a poodle and retrievers are the only dogs in the park. The answer was in fact keyed as *false.* However, the answer was not so obvious for people in the German sample, who were more likely to answer *uncertain* on this item. Ellis researched this issue and found that the poodle originated in Germany as a waterfowl retriever and many Germans still classify the poodle according to its original function. Thus the differences in item scores were probably due to differential knowledge.

Words and their dictionary equivalents in another language may differ in frequency of use and therefore in familiarity. Such discrepancies have a direct influence on the difficulty of the items. However, word frequency dictionaries and familiarity norms can serve a useful purpose in the verification or control of word familiarity in cross-linguistic research. Several sources of relevant data are available for English (see Bradshaw, 1984;

Brown, 1976; Proctor & Vu, 1999) and for French (see Desrochers & Saint-Aubin, 2003).

Problems with experiential equivalence can also occur when the test material involves pictures of objects rather than linguistic stimuli. Picture-naming tests are often used in the assessment of language or communicative disorders. Take the picture of the acorn in the Boston Naming Test, which was developed in the United States to assess confrontational naming of familiar objects (Kaplan, Goodglass, & Weintraub, 1983). How can a stimulus as innocuous as a picture of an acorn cause trouble? It can happen when oak trees are not indigenous to the respondents' environment. This observation was made in the second phase of the Canadian Study of Health and Aging when we found that the acorn item was missed by 90% of our Newfoundland respondents. In checking this problem, we learned that although a few oak trees grow in parks in Newfoundland they are not indigenous and thus may be unfamiliar to many Newfoundlanders. This item was therefore not valid for Newfoundland seniors, as it did not properly measure their ability to name objects. A straightforward strategy for equating picture familiarity is to gather data on naming accuracy across language groups (see Alario & Ferrand, 1999; Cyscowicz, Friedman, Rothstein, & Snodgrass, 1997; Snodgrass & Vanderwart, 1980).

Recommendations for Translating and Adapting Tests

We have reviewed some of the pitfalls that have plagued instrument translation in cross-linguistic research. We shall now address various strategies for enhancing the quality of instrument translation or adaptation across languages. Let us state at the outset that researchers are urged to go beyond literal translation and back translation. Until recently, experts assumed that back translation[2] could uncover any important differences in meaning between the original version of the instrument and its translation (Behling & Law, 2000). However, direct translation and back translation can deal with literal meaning only; they cannot guarantee the general equivalence of the original item and its translation. Even though the criterion of lexical mapping may be met, the focal theme or concept may be modified or lost in the translation process (Hambleton, 2002). Back translation cannot detect differences in conceptual understanding of the question, and so cannot ensure psychological equivalence of the items in a scale or questionnaire (Behling & Law; Hambleton & Patsula, 1998). For example, although the translation of the "poodle" item

[2] In the back-translation design, a test is translated into the target language and then back translated into the source language.

from the CAPS was correct in its literal sense, the intent of the item was lost because people in two language groups had a different understanding of what the item meant. Most experts now prefer the term "adaptation" to "translation" when referring to the process of developing measurement instruments across languages (Geisinger, 1994). Translation is now taken as one of several steps in the process of ensuring that such instruments are meaningfully used in both languages. Detailed guidelines are provided by Geisinger, by Hambleton, and by Vallerand (1989). Banville and her colleagues (2000) summarize Vallerand's methodology and also provide an example from an application of his methods. We shall now summarize the main steps.

Step 1: Verification of Focal Concept Relevance

One must pay close attention to the focal theme or concept of the measurement instrument, determine whether or not it is relevant to both cultures of interest, and, if it is relevant, learn how it is manifested. Hambleton and Patsula (1998) and Behling and Law (2000) describe several relevant strategies, including ethnographic research through observation, interviews, and extended interaction with both cultures to ascertain the relevance of focal concepts.

If the psychological concept of interest is found to be irrelevant to a particular culture, then one should either abandon cross-cultural research on this concept or look for a similar concept that is relevant. Most likely, however, it will be relevant, and researchers then have the option of (a) simultaneously developing their instruments for each language group (Hunt, 1998) — for example, by using the Combined Etic-Emic approach (Hui & Triandis, 1985); (b) developing their instruments for the target language group only (Hambleton, 2002); or (c) adapting an existing source-language instrument for use in the target language (Banville et al., 2000; Bracken & Barona, 1991; Hambleton). Adapting a validated and well-tested instrument for the target language is often the preferred option, when applicable, because of its efficiency and reduced cost.

Step 2: Translation of the Instrument and Development of Preliminary Versions

Hambleton (2002) recommends that professional translators be brought onto the research team; these translators must have an intimate knowledge of the languages and cultures of interest (Geisinger, 1994; Hambleton). Sensitivity to nuances in meaning and expression and awareness of different cultural knowledge and experience can serve to prevent serious problems in item construction. Translators should also be familiar with the concept of interest, the objectives of the measurement

instrument, and the purpose of each item (Geisinger). Therefore, it is important that they be integrated into the research team and collaborate with the researchers to ensure that the meaning is preserved in translation (Behling & Law, 2000).

Once the expertise is in place, double translation (Hambleton, 2002) or double translation/back translation (Vallerand, 1989) can be undertaken. In double translation, two translators each produce an independent version of the measurement instrument in the target language. The two translations are checked for differences, which in turn are resolved through discussions with a third expert (Hambleton) or with the team (Banville et al., 2000). One common version is then developed in the target language (Banville et al.). The double translation/back translation method involves an additional step: two new translators each use the common version to translate the instrument back into the source language (Banville et al.; Vallerand). The translated (or translated and back-translated instrument) is then evaluated by an expert committee (see below). Note that the purpose of these procedures is not to carry out a literal translation but to ensure that the meanings of the original items are retained in the translated instrument (Banville et al.). These procedures can serve to prevent many problems or biases associated with the exclusive use of simple direct translation and back-translation (Vallerand). At this stage of the process, potentially confounding factors such as word complexity, differential experience, and item familiarity in the two languages must be considered, as they can have a considerable effect on respondents' behaviour. These factors sometimes operate in subtle ways. Consider the following example. The Verbal Associative Fluency Scale (FAS; Benton, 1968) is a well-known English verbal fluency test in which respondents are given 1 minute to name all the words they can that begin with the letter *F*. The procedure is then repeated for the letters *A* and *S*. A direct translation of this test into French would be problematic because the number of words that begin with *F, A,* and *S* are different in English and French. All other factors being equal (e.g., word frequency), a smaller sample of suitable words for any of the stimuli would make the task more difficult. The neuropsychologist who adapted this test therefore chose letters with an approximately equivalent number of dictionary words as the original letters; these letters are *T, N,* and *P* (B. Ska, personal communication, April 24, 2003). Resolving this adaptation problem required not only an understanding of the focal concept, but also some knowledge of the determinants of verbal fluency. This example reminds us that expertise in translation and cultural issues, while necessary, is not sufficient to overcome all potential hurdles in cross-linguistic or cross-cultural research.

Step 3: Committee Review and Evaluation of the Preliminary Version

After the preliminary version(s) is/are developed, a committee of experts in the content area, the translators, and the researchers should review and evaluate the translated instrument. Their task is to determine whether the translation is meaningful for all groups and whether its meaning corresponds to the original intent of the items (Banville et al., 2000; Bracken & Barona, 1991; Geisinger, 1994; Hambleton, 2002; Vallerand, 1989). If the measure is to be taken by an interviewer, it may be useful to include bilingual interviewers in this process. Interviewers can provide invaluable insight, as they know about the language used by ordinary people in their area. After difficulties have been resolved, a pre-test version of the instrument is developed.

Step 4: Pre-testing the Instrument

The adapted instrument should be pre-tested on a small sample of people who are representative of the eventual sample. The purpose of the pre-test is not only to collect responses to the items, but also to obtain respondents' feedback on the acceptability and comprehensibility of the items (Vallerand, 1989). Respondents may be interviewed to learn about their interpretation of items; or they may be asked to comment on the clarity of item wording (Vallerand). It is our opinion that focus groups can also be useful at this stage in the development of an instrument. These procedures are all intended to identify problematic items, which should then be revised.

Step 5: Pilot Testing the Instrument

The draft adaptation should be pilot tested to establish its reliability, validity, and acceptability (Banville et al., 2000; Bracken & Barona, 1991; Hambleton, 2002). This pilot testing should probably be done in two phases.

In the first phase, the "test-retest by bilingual subject" procedure developed by Haccoun and recommended by Vallerand (1989) is particularly relevant, because it covers concurrent validity as well as reliability. A group of bilingual participants (Banville et al., 2000, used 20 people) are asked to complete both versions of the instrument and 1 month later are asked to complete them again. The correlation between the original and translated versions assesses concurrent validity, and the test-retest correlation with the same language version assesses reliability (Banville et al.). We recommend that this pilot test be followed up with a larger pilot test of the type described by Bracken and Barona (1991) and Hambleton (2002). The sample for this stage of testing should be large enough (at least 100) to allow for formal factor and item analyses and thereby guide

the next and, it will be hoped, final revision of the instrument. Procedures for testing the conceptual and item equivalence of the instrument for both language groups are described by Behling and Law (2000), Ellis (1989), Hambleton, and Zumbo (in press). Once the test has been refined, it will be necessary for the researchers to validate the assessment in the new language, establish norms with representative samples of respondents, and develop procedures for comparing scores across the two language groups.

Conclusion

Cross-cultural studies may be designed to address theoretical or practical issues. For instance, researchers may want to test a hypothesis regarding how a particular psychological state (e.g., depression) is manifested in different cultures. Alternatively, they may simply want to compare two language groups' satisfaction with the accessibility of medical services. No matter what the focal concepts are, the process of translating measurement items from one language into another always involves potential threats to the validity of the instruments. In this paper, we have considered some of the problems researchers are likely to encounter in adapting a measurement instrument from one language to another. The solutions to these problems are sometimes easy to implement. We have presented a set of guidelines for the development of measurement instruments and their adaptation for cross-linguistic or cross-cultural research. The implementation of these guidelines may be labour-intensive and costly, but it generally is necessary for the construction of reliable and valid measurement instruments.

References

Alario, F. X., & Ferrand, L. (1999). A set of 400 pictures standardized for French: Norms for name agreement, image agreement, familiarity, visual complexity, image variability, and age of acquisition. *Behavior Research Methods, Instruments, and Computers*, *31*, 531–552.

Banville, D., Desrosiers, P., & Genet-Volet, Y. (2000). Translating questionnaires and inventories using a cross-cultural translation technique. *Journal of Teaching in Physical Education*, *19*, 374–387.

Behling, O., & Law, K. (2000). *Translating questionnaires and other research instruments: Problems and solutions*. Sage University Papers Series on Quantitative Applications in the Social Sciences, 07-131. Thousand Oaks, CA: Sage.

Benton, O. (1968). Differential behavioural effects in frontal lobe disease. *Neuropsychologia*, *6*, 53–60.

Bhatnagar, K., & Frank, J. (1997). Psychiatric disorders in elderly from the Indian sub-continent living in Bradford. *International Journal of Geriatric Psychiatry, 12*, 907–912.

Björner, J. B., Kreiner, S., Ware, J. E., Damsgaard, M. T., & Bech, P. (1998). Differential item functioning of the Danish translation of the SF-36. *Journal of Clinical Epidemiology, 51*, 1189–1202.

Bracken, B. A. (1990). Multinational validation of the Spanish Bracken Basic Concept Scale for Cross-Cultural Assessments. *Journal of School Psychology, 28*, 325–341.

Bracken, B. A., & Barona, A. (1991). State of the art procedures for translating, validating and using psychoeducational tests in cross-cultural assessment. *School Psychology International, 12*, 119–132.

Bradshaw, J. L. (1984). A guide to norms, ratings, and lists. *Memory and Cognition, 12*, 202–206.

Brown, A. S. (1976). Catalog of scaled verbal material. *Memory and Cognition, 4*(1B), 1S–45S.

Camilli, G., & Shepard, L. A. (1994). *Methods for identifying biased test items*. Thousand Oaks, CA, London, and New Delhi: Sage.

Crowne, D. P., & Marlow, D. (1960). A new scale of social desirability independent of psychopathology. *Journal of Consulting Psychology, 24,* 349–354.

Curyto, K., Chapleski, E., Lichtenberg, P., Hodges, E., Kaczynski, R., & Sobeck, J. (1998). Prevalence of depression in American Indian elderly. *Clinical Gerontologist, 18,* 19–37.

Cyscowicz, Y. M., Friedman, D., Rothstein, M., & Snodgrass, J. G. (1997). Picture naming by young children: Norms for name agreement, familiarity, and visual complexity. *Journal of Experimental Child Psychology, 65,* 171–237.

Desrochers, A., & Saint-Aubin, J. (2003). *Sources de matériel en français pour l'élaboration des tests de compétences langagières en éducation*. Manuscript submitted for publication.

Edwards, N. (1995). *Predictors of infant-care behaviours among postnatal immigrants*. Unpublished doctoral dissertation, McGill University, Montreal.

Ellis, B. B. (1989). Differential item functioning: Implications for test translations. *Journal of Applied Psychology, 74,* 912–921.

Folstein, M. F., Folstein, S. E., & McHugh, P. R. (1975). Mini-mental State: A practical method for grading the cognitive state of patients for the clinician. *Journal of Psychiatric Research, 12,* 189–198.

Garcia, L. J., & Desrochers, A. (1997). L'évaluation des troubles du langage et de la parole chez l'adulte francophone [Assessment of language and speech disorders in francophone adults]. *Revue d'orthophonie et d'audiologie, 21,* 271–293.

Geisinger, K. F. (1994). Cross-cultural normative assessment: Translation and adaptation issues influencing the normative interpretation of assessment instruments. *Psychological Assessment, 6,* 304–312.

Guarnaccia, P. J., Angel, R., & Worobrey, J. L. (1989). The factor structure of the CES-D in the Hispanic Health and Nutrition Examination Survey – the influences of ethnicity, gender, and language. *Social Sciences and Medicine, 29,* 85–94.

Hambleton, R. K. (2002). Adapting achievement tests into multiple languages for international assessments. In A. C. Porter & A. Gamoran (Eds.), *Methodological*

advances in cross-national surveys of educational achievement (pp. 58–76). Washington: National Academy of Sciences.

Hambleton, R. K., & Patsula, L. (1998). Adapting tests for use in multiple languages and cultures (special issue on Validity Theory and the Methods Used in Validation, edited by B. D. Zumbo). *Social Indicators Research, 45,* 153–171.

Hui, C. H., & Triandis, H. C. (1985). Measurement in cross-cultural psychology: A review and comparison of strategies. *Journal of Cross-Cultural Psychology, 16,* 131–152.

Hunt, S. M. (1998). Cross-cultural issues in the use of quality of life measures in randomized controlled trials. In M. J. Staquet, R. D. Hays, & P. M. Fayers (Eds.), *Quality of life assessment in clinical trials: Methods and practice.* Oxford: Oxford University Press.

Kaplan, E., Goodglass, H., & Weintraub, S. (1983). *Boston naming test.* Philadelphia: Lea & Febiger.

Marshall, S. C., Mungas, D., Weldon, M., Reed, B., & Haan, M. (1997). Differential item functioning in English- and Spanish-speaking older adults. *Psychology and Aging, 12,* 718–725.

McDowell, I. (1994). Canadian study of health and aging: Study methods and prevalence of dementia. *Canadian Medical Association Journal, 150,* 899–913.

Proctor, R. W., & Vu, K.-P. L. (1999). Index of norms and ratings published in the Psychonomic Society journals. *Behavior Research Methods, Instruments, and Computers, 31,* 659–667.

Retief, A. (1988). *Method and theory in cross-cultural psychological assessment.* Pretoria: Human Sciences Research Council.

Sirecci, S. (1997). Problems and issues in linking assessments across cultures. *Educational Measurement: Issues and Practices, 16,* 12–18.

Snodgrass, J. G., & Vanderwart, M. (1980). A standardized set of 260 pictures: Norms for name agreement, image agreement, familiarity, and visual complexity. *Behavior Research Methods, Instruments, and Computers, 28,* 516–536.

Teng, E. L. (1996). *Cross-cultural testing and the Cognitive Abilities Screening Instrument.* In G. Yeo & D. Gallagher-Thomson (Eds.), *Ethnicity and the dementias.* Washington: Taylor & Francis.

Teng, E. L., & Chui, H. C. (1987). The Modified Mini-mental State (3MS) examination. *Journal of Clinical Psychiatry, 48,* 314–318.

Vallerand, R. (1989). Vers une méthodologie de la validation trans-culturelle de questionnaires psychologiques : implications pour la recherche en langue française. *Psychologie canadienne, 30,* 662–678.

Van de Vijver, F. J. R., & Poortinga, Y. H. (1991). *Testing across cultures.* In R. K. Hambleton & J. N. Zaal (Eds.), *Advances in educational and psychological testing: Theory and applications* (pp. 277–308). Evaluation in Education and Human Services series. Boston: Kluwer Academic.

Ware, J. J., & Sherbourne, C. D. (1992). The MOS 36-item short-form health survey (SF-36). I: Conceptual framework and item selection. *Medical Care, 30,* 473–483.

Zumbo, B. D. (in press). Does item-level DIF manifest itself in scale-level analyses? Implications for translating language tests. *Language Testing.*

Authors' Note

Correspondence regarding this paper may be sent to Elizabeth A. Kristjansson, School of Psychology, University of Ottawa, 125 University, Room 415A, Ottawa, Ontario K1N 6N5 Canada.

Elizabeth A. Kristjansson, PhD, is Assistant Professor, School of Psychology, University of Ottawa, Ontario, Canada, and Affiliate Scientist, Centre for Global Health, Institute of Population Health, University of Ottawa. Alain Desrochers, PhD, is Associate Professor, School of Psychology, University of Ottawa. Bruno Zumbo, PhD, is Professor, Faculty of Education, University of British Columbia, Vancouver, Canada.