

Lost in Translation: Methodological Considerations in Cross-Cultural Research

Elizabeth D. Peña

The University of Texas at Austin

In cross-cultural child development research there is often a need to translate instruments and instructions to languages other than English. Typically, the translation process focuses on ensuring linguistic equivalence. However, establishment of linguistic equivalence through translation techniques is often not sufficient to guard against validity threats. In addition to linguistic equivalence, functional equivalence, cultural equivalence, and metric equivalence are factors that need to be considered when research methods are translated to other languages. This article first examines cross-cultural threats to validity in research. Next, each of the preceding factors is illustrated with examples from the literature. Finally, suggestions for incorporating each factor into research studies of child development are given.

In the study of child development, cross-cultural (and intracultural) studies of knowledge acquisition are important for both theoretical and practical reasons. First, cross-cultural methods allow researchers to test, modify, and extend current theories of development (Devescovi & D'Amico, 2005; Katzir, Shaul, Breznitz, & Wolf, 2004; Slobin, 1985). Such research provides insights about the interaction among universal and specific factors in development in the context of social and linguistic variation. Second, the non-English-speaking child population in the United States has increased significantly. According to the U.S. census (Shin & Bruno, 2003), in 1990 14% of the school-age population spoke a language other than English at home, but the proportion had increased to 18% by 2000. There is a practical need to learn about developmental trajectories for children who speak languages other than English or who speak English as a second language. Demographic shifts in the child population due to immigration are paralleled in other countries as well: England (Coleman & Rowthorn, 2004), Italy (Livi Bacci, 2004), Australia (Shah & Long, 2003), and Sweden, Norway, and Denmark, (Kemnitz, 2003). Both theoretical and practical needs drive the heightened interest in cross-cultural research.

Cross-cultural research often necessitates translation of methods (i.e., instruments and instructions to participants) from English to other languages. In the United States, instruments (such as cognitive tests,

language tests, social behavior scales, and school adjustment scales) are typically standardized in English with mainstream American children but they are unlikely to be standardized for other language groups. In addition, question sets, procedures, and coding schemes are developed to address a particular research question. These too are typically developed for mainstream English-speaking children based on what is known about their development. Translation of these instruments and procedures presents particular methodological challenges that can threaten the validity of results. Researchers in child development thus need to be conscious of the pitfalls in translating methods developed for one population and language community to another.

Objectivity is a hallmark in research methodology. However, Greenfield (1994) points out that when psychologists, for example, study development within their own culture, they use their own implicit knowledge of the culture—often unacknowledged—when doing research. This insider's perspective, as practiced by members of a discipline, often becomes the basis for norms (Rogler, 1999), setting the standard for what is studied and how it is studied (Zuckerman, 1988). However, methodological norms developed within and for a given population cannot necessarily be transported without adaptation for the target population.

Development of instrumentation and elicitation procedures appropriate for a question under study is fundamental in research design. Accordingly, detailed methods allow readers to determine the validity and reliability of reported results. When research involves populations that do not speak the

The work for this article was initiated while the author was a Fellow at the Center for Advanced Study in the Behavioral Sciences, Stanford, CA.

Correspondence concerning this article should be addressed to Elizabeth Peña, Department of Communication Sciences and Disorders, University of Texas at Austin, One University Station A1100, Austin, TX 78712. Electronic mail may be sent to lizp@mail.utexas.edu.

majority language (e.g., English in the United States), particular attention to development of instrumentation and procedures is needed to ensure validity and reliability. For instance, the *Publication Manual of the American Psychological Association (2001)* has a section on reduction of language bias. Guidelines state that the language used in the procedure of a study needs to be specified. Furthermore, the method used to translate test instruments to a language other than English must be detailed. But mere translation of elicitation procedures and instrumentation is not sufficient to guard against potential cultural bias and therefore validity threats.

Challenges to Validity in Translation

Bias is a distinct threat to validity in translation of methods in cross-cultural research. The literature on bias in test development provides a useful framework for discussion of linguistically appropriate research methods. An important principle for such a discussion is the notion of fairness in test development. Fairness is evaluated in the context of the goals or function of the test instrument. Definitions of fairness include equal treatment in context and purpose of testing, and comparable opportunity to demonstrate abilities on the construct the test is intended to measure (*Standards for Educational and Psychological Testing, 1999*). An important methodological goal therefore is to ensure equivalence at the level of context and opportunity when one is designing cross-cultural research studies of child development. Equivalence may be at the level of stimuli for the purpose of exploring similarity and variation in response, or equivalence may be at the level of outcome for the purpose of understanding differences in circumstances that bring about specific developmental results.

The principles of equal treatment and comparable opportunity can be applied to development of cross-cultural methods. Instructions and tests used across languages need to be equivalent to provide equal opportunity to demonstrate the skill under study. However, translation may not by itself ensure equal opportunity for the participants to demonstrate their abilities. Instructions to participants and the content of the instrument(s) used to gather data are potential sources of bias. In addition to linguistic equivalence, the notion of equivalence can be interpreted and applied in several additional ways: functional equivalence, cultural equivalence, and metric equivalence (Arnold & Matus, 2000; Bracken & Barona, 1991; Erkut, Alarcón, Coll, Tropp, & García, 1999; Geisinger, 1994; Rogers, Gierl, Tardif, Lin, & Rinaldi, 2003;

Rogler, 1999; Sechrest, Fay, & Hafeez Zaidi, 1972; Sireci & Berberoglu, 2000; *Standards for Educational and Psychological Testing, 1999*; Valencia & Rankin, 1985). The type of equivalence identified as necessary depends on a study's goals and involves consideration of stimuli and outcomes. Here, principles drawn from disciplines that have a long tradition of cross-cultural research, such as anthropology and sociology, as well as applied fields, such as clinical psychology and nursing, guide development of a framework appropriate for research in child development.

Linguistic equivalence typically refers to translating instructions and instruments, and checking the translation with methods such as back-translation (translation from the first language to the second, and then back to the first by a second person; *Standards for Educational and Psychological Testing, 1999*) or expert review (Hambleton, 2001). Functional equivalence means that the instructions and instrument will elicit the same target behavior (Greenfield et al., 2006). Cultural equivalence considers how respondents will interpret a given direction or test item and develops items that tap the same cultural meaning for each cultural linguistic group (Alonso et al., 1998). Metric equivalence has to do with the difficulty of the specific item expressed in two distinct languages (Azen et al., 1999; Kim, Han, & Phillips, 2003; Muñiz, Hambleton, & Xing, 2001) and is essential for development of ability tests for example. Together, these types of equivalence provide a way to examine potential methodological bias. Examples from the literature demonstrate the challenges to achievement in each of these kinds of equivalence.

Linguistic Equivalence

Direct translation usually satisfies the standard for ensuring linguistic equivalence. Researchers employ two main types of techniques when translating instruments and instructions. In translation and back-translation (Arnold & Matus, 2000; Beck, Bernal, & Froman, 2003; Brislin, 1986; Hambleton, 2001; Rogers et al., 2003) a translator first translates the instrument or instructions from the source language to the target language. A second translator then independently translates the target version back to the source language. The original and back-translated versions are compared to identify differences, which are then resolved. This procedure is akin to the game of "telephone" but with a cross-linguistic twist. Another technique is to have a native language speaker review the translation to ensure its accuracy. The main goal for linguistic equivalence is to make certain that the words and linguistic meaning used in

the instruments and instructions are the same for both versions (Grisay, 2003; Sireci & Berberoglu, 2000).

A problem with linguistic equivalence is that even if words are the same across two sets of methods there are potential differences that may result in different patterns of responses. That is, the same stimuli may result in different outcomes. These different patterns of response may be due to differences in cultural interpretation, familiarity, or frequency of occurrence. If these are the research questions, linguistic equivalence is sufficient and appropriate. If, however, the purpose of a translated instrument is to make a judgment of a developmental status, linguistic equivalence without consideration of functional, cultural, and metric equivalence may introduce bias. For example, when the Preschool Language Scale-3 (Zimmerman, Steiner, & Pond, 1992), a test of linguistic and conceptual development, was initially translated to Spanish (Zimmerman, Steiner, & Pond, 1993) all the test items were retained in the same order. An item analysis by Restrepo and Silverman (2001) demonstrated that although the items were linguistically equivalent, item difficulty at each age level was not. All concepts and linguistic forms are not learned at the same point in development in all languages; thus, these may be easier or harder cross-linguistically. For example, understanding object functions was more difficult in English than in Spanish, but prepositions were more difficult in Spanish than in English. Use of such an instrument in a comparative study of linguistic or cognitive development could be misleading because it possibly under- or over-estimates developmental status. The most recent Spanish version (Zimmerman, Steiner, & Pond, 2002) used many of the same (English) items, introduced new items based on milestones of Spanish language development (e.g., gender agreement), and based item order on Spanish difficulty to yield psychometrically equivalent tests (discussed further later). Such tests are more appropriate for evaluation of development because they compare an individual child against a linguistically and culturally appropriate standard.

Another example that illustrates validity threats in translation concerns the ways target behaviors are elicited. The adaptation of the Peabody Picture Vocabulary Test-Revised (PPVT-R; Dunn & Dunn, 1981) to the Spanish Test de Vocabulario en Imagenes Peabody (TVIP, Dunn, Padilla, Lugo, & Dunn, 1986) provides such a case. In Spanish, as in English, this test is a single-word recognition task. Children hear a word (usually a noun) spoken by the examiner and select one of four pictures that best depicts the given word. In both versions the words are presented without an article (e.g., "dog" not "the dog"). In English, the article carries

little linguistic information and nouns frequently occur without articles. In Spanish however, nouns are typically accompanied by the article, which marks gender and number (*la* – feminine singular, *las* – feminine plural, *el* – masculine singular, *los* – masculine plural). The test manual of the TVIP, therefore, instructs examiners not to include the article when saying the word because it may provide children with additional cues that will enable them to select the correct word on the basis of the gender and number information (for an experimental example of how children use grammatical gender in Spanish, see Lew-Williams & Fernald, in press). Spanish-speaking examiners often object to this instruction, calling it "unnatural Spanish." Omitting the article could result in a functional difference unintentionally affecting test performance because Spanish-speaking children do not typically hear nouns without their articles. An alternative way to have constructed the test, allowing the more typical use of the article + noun, would have been to control for gender and number. That is, each of the four alternates (the target and foils) could have been of the same gender and number so as not to provide an extra cue yet letting children hear the target word in its familiar context.

These two examples show that establishing linguistic equivalence through the established methods of expert translation and back-translation in instrumentation and instructions or elicitation of targets is not always sufficient for development of methods in studies of groups who do not speak English. Although the examples target test instruments specifically, they illustrate the threats that are inherent in using translated methods in cross-cultural studies of development. Instructions or elicitation procedures must also be scrutinized to ensure equal opportunity to demonstrate the target ability. Examination of functional, cultural, and metric equivalence may be needed to guard against validity threats.

Functional Equivalence

Rogler (1999) argued that preservation of the language used in the original language version—or linguistic equivalence—is a potential source of cultural insensitivity if the translation yields functional differences. In other words, translation from one language to another can result in incongruity in meaning, threatening content validity of a study's methods. Functional equivalence addresses some of these threats by ensuring that the instrument and elicitation method allow examination of the same construct. This aspect of translation is often overlooked in favor of achieving uniformity in instrumentation and procedures.

One translation method with the purpose of equalizing concepts or function over linguistic equivalence is referred to as “decentering” (Sechrest et al., 1972). Decentering is often used by professional translators to obtain equivalence in meaning and salience with respect to the respondent, in combination with the translation/back-translation approach. This procedure may yield an instrument with translated items that have shifted away from the source instrument’s wording to represent the concept in a linguistically familiar way in the target language.

Another translation method is a “dual-focus” approach (Erkut et al., 1999). This method uses a research team drawn from both of the cultural and linguistic groups under study. The instrumentation and instructions to be used in the research study are developed simultaneously in the two languages so that methods that are linguistically appropriate for each of the target groups focus on equality in clarity (rather than linguistic equivalence) for each. Thus, instruments are parallel with respect to the behavior or concepts tested but with different stimuli.

The following example demonstrates how functional equivalence in instrumentation and procedures can be obtained by use of a combination of translation, back-translation, decentering, and dual-focus procedures (see Bedore, Peña, García, & Cortez, 2005). The Bilingual English Spanish Assessment (BESA; Peña, Gutierrez-Clellen, Iglesias, Goldstein, & Bedore, 2007) is designed to identify language impairment in Latino children between the ages of 4 years 0 months and 6 years 11 months in the United States. The BESA contains two language versions (English and Spanish) targeting four domains: semantics, morpho-syntax, pragmatics, and phonology.

Development of the characteristic properties items from the semantics subtest of the BESA provides an illustration of functional equivalence. These items were designed to elicit descriptions of common objects (e.g., school bus, truck, spoon, and fork) from children ages 4, 5, and 6. Of interest here are the question frames used to elicit description in each of the two languages. Observation of a bilingual pre-school classroom teacher–student interaction was the beginning point for development of question frames and indicated that different types of questions were used in the two languages. During item development, the different question types identified in the classroom were piloted with a small number of children and included Spanish and English versions (using both dual-focus and back-translation techniques) of “tell me about . . .,” “describe . . .,” “tell me . . .,” and “tell me three things about . . .,” among others. The pilot data indicated that “tell me three things about . . .”

in English and “describe . . .” in Spanish yielded functionally equivalent language performance (resulting in decentered elicitation frames). The different question forms are appropriate for each language and have the effect of eliciting similar target behavior in each of the two languages. These question frames were incorporated into the BESA and tested with a larger number of children (Bedore et al., 2005). Statistical analyses indicated similar performance in each language for both monolingual and bilingual children. Thus elicitation frames for each language in the final version are functionally equivalent and linguistically different, yet both elicit linguistically similar responses. Attention to functional equivalence allows children to demonstrate their knowledge in an elicitation context that is familiar in each language (Fagundes, Haynes, Haak, & Moran, 1998; Peña, 2001). This type of equivalence levels the cross-cultural playing field.

Cultural Equivalence

In their study, van der Veer, Ommundsen, Hak, and Larsen (2003) pointed out that items may have different salience for different cultural and linguistic groups, even if the items meet the criteria for linguistic and functional equivalence. Disparities in salience may be due to the distinct cultural and historical ways in which concepts are interpreted by respondents. Cultural equivalence with respect to respondents’ interpretations and responses to given items needs to be explored when one is developing methods and procedures.

The notion of cultural equivalence is related to that of functional equivalence (Arnold & Matus, 2000; Geisinger, 1994; Muñoz et al., 2001; Sechrest et al., 1972). Cultural equivalence focuses more centrally on the way members of different cultural and linguistic groups view or interpret the underlying meaning of an item. Cultural interpretations may affect the ways individuals respond to instructions and research instruments, including standardized and nonstandardized tests (Canino & Guarnaccia, 1997; Hendrickson, 2003).

Culturally determined definitions of developmental abilities such as knowledge (Zambrano & Greenfield, 2004), creativity (Baldwin, 2001), and language (McCollum & Chen, 2001; Posada, Carbonell, Alzate, & Plata, 2004; Suizzo, 2004; Vigil, 2002) may also affect the ways children and families from linguistically and culturally diverse backgrounds report information. Zambrano and Greenfield (2004) hypothesized that “different ethnic groups have their own implicit, informal theories of knowledge and that these ethno-theories form the assumptions on which the

explicit formal theories are based" (p. 251). That is, Western theories of intelligence rest on cultural-specific assumptions (see also Greenfield, 1997). Zambrano and Greenfield illustrated that the concept of knowledge has core understandings, albeit with overlapped meanings in American and in several Maya (Tzotzil-speaking) communities. The American English word *know* refers to facts, theories, and practice or "know-how." The Tzotzil word *na* translates to the English *know* but in addition to facts, theory, and practice, it also implies habitual practice that indicates mastery and is part of the person's character.

This critical distinction of habitual practice or mastery is similar to an example reported by Peña and Jackson (2000), in which a Mexican immigrant child about 2-1/2 years old was referred for a speech-language evaluation because the developmental milestone of first words was reported by the mother to have been reached at 24 months of age, a significant delay. In the initial interview, the Spanish-speaking speech-language pathologist asked the mother for examples of the child's first words. All the examples were of word combinations, rather than single words, clearly within the age-expected range. In this example, the mother's ethnotheory of "learning first words" was talking to communicate using word combinations. Earlier use of single words did not count because they were not evidence of mastery of talking. Simply asking when the child first began to talk did not produce a response that had the same cultural meaning it might have for a mainstream, English-speaking, American mother.

A study by Garstein, Slobodskaya, and Kinsht (2003) provides another illustration of potential culturally biased responses. In this study of infant temperament, Russian and American mothers were asked to complete the Infant Behavior Questionnaire-Revised (IBQ-R; Gartstein & Rothbart, 2003), which consists of 14 scales—for example, activity level, smile and laughter, soothability, sadness, and vocal reactivity. The materials were translated to Russian with a translation and back-translation technique. Results comparing Russian and American infants confirmed expected cultural differences. Specifically, U.S. mothers reported "higher levels of smiling and laughter, high and low intensity pleasure, perceptual sensitivity and vocal reactivity" (p. 322). On the other hand, Russian mothers reported higher distress to limitations. The authors of this study acknowledged that the emphasis on development of certain behaviors with respect to expression of emotions may be directly linked to the types of emotions that their children then demonstrated. However, there may be another, related factor at work—that

mothers interpret their babies' temperamental characteristics on the basis of their cultural expectations. Data on bilingual Russian-English personal narratives lend credibility to this argument. Marian and Kaushanskaya (2004) compared autobiographical memories of bilingual Russian-English speakers in each of their two languages to explore the notion of self-construal. They found that narratives retrieved in Russian included fewer personal pronouns and more group pronouns (an indicator of collectivism) than those retrieved in English (indicating individualism). Furthermore, emotional intensity of the narratives had different patterns in each of the two languages. Memories encoded in Russian were less positive than those encoded in English.

A way to test the possibility of cultural bias and to disentangle actual infant behaviors from a mother's interpretation of those behaviors would be to conduct a study of item salience, similar to that described by van der Veer et al. (2004). Applied to the study under discussion, a subset of Russian and American mothers would rate the temperament of a set of control infants via videotape using the IBQ-R. Comparisons would then be made for the two samples. If there were systematic differences in ratings by nationality, they would point to cultural differences in how mothers interpret the same behavior. These ratings could then be used to calibrate the responses from the larger sample to examine actual differences in infant behaviors independently of cultural bias. Furthermore, these comparisons could be used to shed light on the nature of such cultural differences.

Metric Equivalence

The final aspect of equivalence discussed here is that of metric equivalence. Metric equivalence refers to equivalence in item or question difficulty. This type of equivalence is particularly important when one is developing instruments in more than one language or adapting an instrument from one language to another. Review of the methods used for adapting and developing vocabulary assessments from English to other languages exemplifies the methodological challenges.

An early example of adaptation of an English test to another language is illustrated by the adaptation of the PPVT-R (Dunn & Dunn, 1981) to a Spanish version, the TVIP (Dunn et al., 1986). The original English corpus was based on English-language dictionaries, and selections were based on item difficulty to sample across a broad age range. These items were translated to Spanish, and similar field testing was conducted in Spain to select items on the basis of item

difficulty. These items compose the TVIP. After standardization in Spain, standardization was completed in Mexico and Puerto Rico with monolingual Spanish speakers. Although the TVIP was deliberately not normed with bilingual U.S. populations, Dunn et al. (1986) provided comparative information drawn from pilot studies of bilingual children. The Mexican and Puerto Rican children performed below the mean on the same item set compared with the Spanish children. In addition, the U.S. bilingual children performed about 1 *SD* below their monolingual Spanish-speaking Mexican and Puerto Rican counterparts on this test.

This adaptation has been criticized on several methodological and psychometric bases (Berliner, 1988; Cummins, 1988; Mercer, 1988; Prewitt Diaz, 1988; Trueba, 1988; Willig, 1988). An important lesson for the current purpose is that different dialects and usages vary across and within languages. First, beginning solely with an English corpus is not appropriate because words may have different frequencies and different uses in the two languages. Second, using the same words for Spanish, Mexican, and Puerto Rican Spanish tests may have resulted in differences in performance because vocabulary use and frequency are different for these three populations. Specifically, names for things differ across dialects of Spanish (e.g., *pantallas*, *aretes*, and *pendientes* refer to "earrings" in Puerto Rican, Mexican, and Castilian Spanish, respectively) as it does for dialects of English (e.g., *torch* and *flashlight* name the same thing in British vs. American English).

Developing parallel vocabulary measures based on word frequency rather than translation may provide a superior way to develop psychometrically parallel instruments. Tamayo (1987) developed an English word list and two Spanish word lists, one matched to English on the basis of translation and the other matched on the basis of item frequency. The English and Spanish lists were administered to two groups of 80 eighth-grade students, one group fluent in English and the other in Spanish. Children responded by providing a definition of each target word. The two groups were further matched by gender, age, and academic achievement. Comparison of the children's performance indicated that the English and Spanish versions matched by frequency yielded comparable performance by the two groups, whereas those matched by translation resulted in group differences. These results imply that metric equivalence as applied to instrument development should consider lexical frequency in each of the target languages if other psychometric data are not available.

The procedures developed by Fenson, Bates, and colleagues for the adaptation of the Bates–MacArthur

Communication Developmental Inventories (CDI; Fenson et al., 1993) illustrate how several methods were used in combination for development of this instrument for various languages including Italian (Caselli et al., 1995), Mexican Spanish (Jackson-Maldonado, Thal, Marchman, Bates, & Gutierrez-Clellen, 1993), Cuban Spanish (Pearson & Fernandez, 1994), Mandarin (Tardif, Gelman, & Xu, 1999), Finnish (Lyytinen, Poikkeus, Leiwo, Ahonen, & Lyytinen, 1996), Canadian French (Poulin-Dubois, Graham, & Sippola, 1995), and Hebrew (Maitel, Dromi, Sagi, & Bornstein, 2000). For each adaptation careful attention was paid to typology of the target language, word frequencies, and word class (Fenson et al., 1994). For example, the Italian version differs from the American English version on the specific content words included, but both versions use an approximately equal number of words from different categories (e.g., animal, food, and clothing items have the same number of items but consist of different words; Caselli et al., 1995; Caselli, Casadio, & Bates, 1999). The grammatical function word categories reflect the structural differences of the two languages in the types of words and number of each type. Specifically, the Italian version includes adverbials but the English version does not; the English version includes 27 prepositions and the Italian version include 17. A section was added in the Italian version to examine verb conjugation and noun declension. For Mexican Spanish, modifications were also made to render the instrument linguistically and culturally relevant (Jackson-Maldonado et al., 1993). As in the Italian version, lexical categories were added to reflect verb conjugation; gender in articles, pronouns, and adjectives; and number in articles and pronouns. As in the examples for Italian, the Spanish content reflected culturally appropriate vocabulary and routines. For example, *tortillitas* and *ojitos* replaced *pat-a-cake* and *peek-a-boo*.

Within-language adaptations were also made on the CDI specific for the target population. For example, Hamilton, Plunkett, and Schafer (2000) adapted the CDI for a British population. Examination of their most recent word list (referred to as the Oxford 1998 CDI) indicates several changes from the American version. For example, the Oxford 1998 CDI includes *pushchair*, *brick*, *biscuit*, *sweets*, and *nappy*, whereas the American version includes *stroller*, *block*, *cookie*, *candy*, and *diaper*. Furthermore, analysis of word frequencies indicated that the items on the Oxford 1998 CDI occurred more frequently in British English than in American English, providing evidence of validity for the target population.

In sum, for these adaptations, the authors used several methods in combination, including deriving

items from studies of natural language samples, available corpora from language experiments, vocabulary lists from other scales in the target language, and vocabulary lists from CDIs already adapted to other languages. Furthermore, they asked informants from the target population to review the words and to identify irrelevant words, as well as to add relevant words that had not been included. The use of these procedures has resulted in instruments that are appropriate for the populations for which they are intended and that have a high degree of reliability. Cross-cultural comparisons using these instruments are likely to yield valid results.

Similarly, in development of the tasks for the BESA Semantics Test, Peña, Bedore, and Rappazzo (2003) compared Spanish- and English-speaking children's performance on six tasks (analogies, characteristic property, categorization, functions, linguistic concepts, and similarities and differences). Each task had 86 items in each language: 12 items per task and 26 items for categorization. Items were developed using a combination of dual focus, where half of the items were developed for each language independently; translation, where each item was translated to the other language (half to English, half to Spanish); and decentering, where each item was further adapted so that the question context was different but relevant for each group. The contextualizing aspect was conducted in part so that children would not recognize the question if tested in both languages (see Peña et al., 2003, for more details). Analyses revealed significant Language \times Task interactions. Some tasks were relatively easier in English than in Spanish (e.g., similarities and differences) but other tasks were easier in Spanish than in English (e.g., functions). For the second stage of data collection, representation of items was based primarily on item difficulty, while attempting to represent all the task types. Thus, the item configuration for each language is different. Finally, items are arranged by difficulty for the target language. This configuration and arrangement results in psychometrically parallel tests that are not linguistically equivalent but that validly assess semantic performance in each language. Preliminary data analysis comparing children with and without language impairment in Spanish and English on these same tasks indicates differential performance by task and (test) language. That is, certain tasks function better to differentiate language impairment in Spanish, whereas other tasks are better suited for English. For example, more functions items discriminated between children with and without language impairment in Spanish, but more similarities-and-differences items discriminated these children in

English. In sum, psychometric adaptation may mean that the resulting instruments will have different items and different arrangement of items.

Guidelines for Translation of Instruments for Cross-Cultural Research

There is no question that cross-cultural research is beneficial and desirable from both theoretical and practical perspectives. Theoretically, cross-cultural research allows researchers to test and extend theories of development. A cross-cultural approach can help to identify universals in development and to discover variation attributable to linguistic and cultural differences. Most of the examples described here are based on tests of language. Nonetheless, the principles apply to other domains of developmental research that require translation from English to another language. Knowledge of how development unfolds in different linguistic and cultural contexts informs application of the state of the art to a broader population of children. Consideration of additional aspects of equivalence, such as functional, cultural, and metric equivalence, can reduce potential methodological bias.

For translation of instructions to participants and of instrumentation, first consider whether bias will be introduced to the study. Techniques such as translation and back-translation result in well-translated instructions and instruments, but they may not provide equally relevant methods for the populations under study. Decentering—that is, adapting the question or item so that it is culturally familiar—can also be used in conjunction with translation.

When instruments and instructions are adapted from well-established protocols into other languages, assuring functional equivalence is essential if the study's focus is on whether children are able to perform a given task. Functional equivalence can be evaluated by interviewing informants, conducting literature reviews, and examining available corpora. The CHILDES (MacWhinney, 2000) and SALT (Miller & Iglesias, 2005) reference databases provide rich sources of raw data that can be analyzed to develop functionally equivalent instructions and instrumentation. Examination of word frequencies and typical question forms, for example, can be used to guide selection of target vocabulary or development of question frames.

Instructions, questions, and tasks will not have the same degree of cultural relevance for participant groups. In particular, when tasks are set up in the same way for different cultural/linguistic groups,

differences in outcomes may be influenced by differences in expectations or interpretation rather than by differences in the trait under study. The potential for cultural mismatch highlights the need to consider cultural equivalence in developing methods for cross-cultural research. Debriefing participants during the pilot stage of an adapted method may also help researchers understand how participants might interpret instructions or questions and can also be used to explore the nature of cultural differences. Asking for examples or explanations is another way to understand research participants' response patterns. These added steps will allow for the examination of potential culturally driven responses to help disentangle true from ostensible differences.

Psychometric equivalence is particularly critical for development of instrumentation if comparisons between groups will be made that focus on judgments of ability. Items may not be equally difficult across languages even if the target concept or question occurs in both languages. Some types of items may be rendered more or less complex when translated; words selected in the translation may have different frequencies of occurrence and influence difficulty. These considerations are especially important in translation of instruments that test linguistic and cognitive ability, but they are also important in academic domains such as math (Abedi & Lord, 2001; Towse & Saxton, 1998).

For development of psychometric instruments, item difficulty needs to be taken into account. The conventional way to determine item difficulty is to calculate the percentage of participants (e.g., at a given age) who respond correctly to an item. Another way to index difficulty is to examine frequency of occurrence in the target language (Tamayo, 1987, 1990) or to examine age of acquisition of target words or concepts (Bates et al., 2003).

In summary, four features that need to be considered when one conducts research across cultural/linguistic groups include: linguistic equivalence, functional equivalence, cultural equivalence, and metric equivalence. Attention to these methodological features is critical for establishing a study's validity. Their consideration and application when needed will reduce validity threats in cross-cultural research.

References

- Abedi, J., & Lord, C. (2001). The language factor in mathematics tests. *Applied Measurement in Education, 14*, 219–234.
- Alonso, J., Black, C., Norregaard, J.-C., Dunn, E., Andersen, T. F., Espallargues, M., et al. (1998). Cross-cultural differences in the reporting of global functional capacity: An example in cataract patients. *Medical Care, 36*, 868–878.
- Arnold, B. R., & Matus, Y. E. (2000). Test translation and cultural equivalence methodologies for use with diverse populations. In I. Cuellar & F. A. Paniagua (Eds.), *Handbook of multicultural mental health* (pp. 121–136). San Diego, CA: Academic Press.
- Azen, S. P., Palmer, J. M., Carlson, M., Mandel, D., Cherry, B. J., Fanchiang, S.-P., et al. (1999). Psychometric properties of a Chinese translation of the SF-36 Health Survey Questionnaire in the well elderly study. *Journal of Aging & Health, 11*, 240–251.
- Baldwin, A. Y. (2001). Understanding the challenge of creativity among African Americans. *Journal of Secondary Gifted Education, 12*(3), 121–125.
- Bates, E., D'Amico, S., Jacobsen, T., Székely, A., Andonova, E., Devescovi, A., et al. (2003). Timed picture naming in seven languages. *Psychonomic Bulletin & Review, 10*, 344–380.
- Beck, C. T., Bernal, H., & Froman, R. D. (2003). Methods to document semantic equivalence of a translated scale. *Research in Nursing & Health, 26*, 64–73.
- Bedore, L. M., Peña, E. D., García, M., & Cortez, C. (2005). Conceptual versus monolingual scoring: When does it make a difference? *Speech, Language, Hearing Services in Schools, 36*, 188–200.
- Berliner, D. C. (1988). Meta-comments: A discussion of critiques of L.M. Dunn's monograph bilingual Hispanic children on the U.S. Mainland. *Hispanic Journal of Behavioral Sciences, 10*, 273–299.
- 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.
- Brislin, R. (1986). Back-translation methods: The wording and translation of research instruments. In W. J. Lonner & J. W. Berry (Eds.), *Field methods in cross-cultural research* (pp. 137–164). Beverly Hills, CA: Sage.
- Canino, G., & Guarnaccia, P. (1997). Methodological challenges in the assessment of Hispanic children and adolescents. *Applied Developmental Sciences, 1*, 124–134.
- Caselli, M. C., Bates, E., Casadio, P., Fenson, J., Fenson, L., Sanderl., et al. (1995). A cross-linguistic study of early lexical development. *Cognitive Development, 10*, 159–199.
- Caselli, M. C., Casadio, P., & Bates, E. (1999). A comparison of the transition from first words to grammar in English and Italian. *Journal of Child Language, 26*, 69–111.
- Coleman, D., & Rowthorn, R. (2004). The economic effects of immigration into the United Kingdom. *Population and Development Review, 30*, 579–624.
- Cummins, J. (1988). "Teachers are not miracle workers": Lloyd Dunn's call for Hispanic activism. *Hispanic Journal of Behavioral Sciences, 10*, 263–272.
- Devescovi, A., & D'Amico, S. (2005). The competition model: Crosslinguistic studies of online processing. In M. Tomasello & D. I. Slobin (Eds.), *Beyond nature–nurture: Essays in honor of Elizabeth Bates* (pp. 165–191). Mahwah, NJ: Erlbaum.

- Dunn, L., & Dunn, L. (1981). *Peabody Picture Vocabulary Test—Revised*. Circle Pines, MN: American Guidance Service.
- Dunn, L., Padilla, R., Lugo, S., & Dunn, L. (1986). *Test de Vocabulario en Imágenes Peabody*. Circle Pines, MN: American Guidance Service.
- Erkut, S., Alarcón, O., Coll, C. G., Tropp, L. R., & García, H. A. V. (1999). The dual-focus approach to creating bilingual measures. *Journal of Cross-Cultural Psychology, 30*, 206–218.
- Fagundes, D., Haynes, W., Haak, N., & Moran, M. (1998). Task variability effects on the language test performance of southern lower socioeconomic class African American and Caucasian five-year-olds. *Language, Speech & Hearing Services in the Schools, 29*, 148–157.
- Fenson, L., Dale, P. S., Reznick, J. S., Bates, E., Thal, D. J., & Pethick, S. J. (1994). Variability in early communicative development. *Monographs of the Society for Research in Child Development, 59*(5, Serial No. 242).
- Fenson, L., Dale, P. S., Reznick, J. S., Thal, D., Bates, E., Hartung, J. P., et al. (1993). *MacArthur Communicative Development Inventories: User's guide and technical manual*. Baltimore: Brookes.
- Gartstein, M. A., & Rothbart, M. K. (2003). Studying infant temperament via the Revised Infant Behavior Questionnaire. *Infant Behavior & Development, 26*, 64–86.
- Gartstein, M. A., Slobodskaya, H. R., & Kinsht, I. A. (2003). Cross-cultural differences in temperament in the first year of life: United States of America (U.S.) and Russia. *International Journal of Behavioral Development, 27*, 316–328.
- Geisinger, K. F. (1994). Cross-cultural normative assessment: Translation and adaptation issues influencing the normative interpretation of assessment instruments. *Psychological Assessment, 6*, 304–312.
- Greenfield, P. M. (1994). Independence and interdependence as developmental scripts: Implications for theory, research, and practice. In P. M. Greenfield & R. R. Cocking (Eds.), *Cross-cultural roots of minority child development* (pp. 1–37). Hillsdale, NJ: Erlbaum.
- Greenfield, P. M. (1997). You can't take it with you: Why ability assessments don't cross cultures. *American Psychologist, 52*, 1115–1124.
- Greenfield, P. M., Trumbull, E., Keller, H., Rothstein-Fisch, C., Suzuki, L., & Quiroz, B. (2006). Cultural conceptions of learning and development. In P. A. Alexander, P. R. Pintrich, & P. H. Winne (Eds.), *Handbook of educational psychology* (2nd ed., pp. 675–694). Mahwah, NJ: Erlbaum.
- Grisay, A. (2003). Translation procedures in OECD/PISA 2000 international assessment. *Language Testing, 20*, 225–240.
- Hambleton, R. K. (2001). The next generation of the ITC test translation and adaptation guidelines. *European Journal of Psychological Assessment, 17*, 164–172.
- Hamilton, A., Plunkett, K., & Schafer, G. (2000). Infant vocabulary development assessed with a British Communicative Development Inventory. *Journal of Child Language, 27*, 689–705.
- Hendrickson, S. G. (2003). Beyond translation... Cultural fit. *Western Journal of Nursing Research, 25*, 593–608.
- Jackson-Maldonado, D., Thal, D., Marchman, V. A., Bates, E., & Gutierrez-Clellen, V. (1993). Early lexical development in Spanish-speaking infants and toddlers. *Journal of Child Language, 20*, 523–549.
- Katzir, T., Shaul, S., Breznitz, Z., & Wolf, M. (2004). The universal and the unique in dyslexia: A cross-linguistic investigation of reading and reading fluency in Hebrew- and English-speaking children with reading disorders. *Reading & Writing, 17*, 739–768.
- Kemnitz, A. (2003). Immigration, unemployment and pensions. *Scandinavian Journal of Economics, 105*, 31–48.
- Kim, M., Han, H.-R., & Phillips, L. (2003). Metric equivalence assessment in cross-cultural research: Using an example of the Center for Epidemiological Studies—Depression Scale. *Journal of Nursing Measurement, 11*, 5–18.
- Lew-Williams, C., & Fernald, A. (2007). Young children learning Spanish make rapid use of grammatical gender in spoken word recognition. *Psychological Science, 18*, 193–198.
- Livi Bacci, M. (2004). The population of the developed countries: Decreasing returns? *Review of Economic Conditions in Italy, January–April (1)*, 27–50.
- Lyytinen, P., Poikkeus, A. M., Leiwo, M., Ahonen, T., & Lyytinen, H. (1996). Parents as informants of their child's vocal and early language development. *Early Child Development and Care, 126*, 15–25.
- MacWhinney, B. (2000). *The CHILDES project: Tools for analyzing talk*. Mahwah, NJ: Erlbaum.
- Maitel, S. L., Dromi, E., Sagi, A., & Bornstein, M. H. (2000). The Hebrew Communicative Development Inventory: Language specific properties and cross-linguistic generalizations. *Journal of Child Language, 27*, 43–67.
- Marian, V., & Kaushanskaya, M. (2004). Self-construal and emotion in bicultural bilinguals. *Journal of Memory & Language, 51*, 190–201.
- McCollum, J. A., & Chen, Y.-J. (2001). Maternal roles and social competence: Parent–infant interactions in two cultures. *Early Child Development and Care, 166*, 119–133.
- Mercer, J. (1988). Ethnic differences in IQ scores: What do they mean? (A response to Lloyd Dunn). *Hispanic Journal of Behavioral Sciences, 10*, 199–218.
- Miller, J., & Iglesias, A. (2005). *Systematic Analysis of Language Transcripts—SALT V9*. University of Wisconsin, Language Analysis Laboratory, Waisman Center.
- Muñoz, J., Hambleton, R. K., & Xing, D. (2001). Small sample studies to detect flaws in item translations. *International Journal of Testing, 1*, 115–135.
- Pearson, B. Z., & Fernandez, S. C. (1994). Patterns of interaction in the lexical growth in two languages of bilingual infants and toddlers. *Language Learning, 44*, 617–653.
- Peña, E. D. (2001). Assessment of semantic knowledge: Use of feedback and clinical interviewing. *Seminars in Speech and Language, 22*, 51–63.
- Peña, E. D., Bedore, L. M., & Rappazzo, C. (2003). Comparison of Spanish, English, and bilingual children's performance across semantic tasks. *Language, Speech & Hearing Services in the Schools, 34*, 5–16.
- Peña, E. D., Gutierrez-Clellen, V. F., Iglesias, A., Goldstein, B., & Bedore, L. M. (2007). *Bilingual English Spanish assessment*. Manuscript in preparation.

- Peña, E. D., & Jackson, J. (2000). The social and cultural bases of communication. In R. Gillam, T. Marquardt, & F. Martin (Eds.), *Communication sciences & disorders: From science to clinical practice* (pp. 63–84). San Diego, CA: Singular.
- Posada, G., Carbonell, O. A., Alzate, G., & Plata, S. J. (2004). Through Colombian lenses: Ethnographic and conventional analyses of maternal care and their associations with secure base behavior. *Developmental Psychology, 40*, 508–518.
- Poulin-Dubois, D., Graham, S., & Sippola, L. (1995). Early lexical development: The contribution of parental labeling and infants categorization abilities. *Journal of Child Language, 22*, 325–343.
- Prewitt Diaz, J. O. (1988). Assessment of Puerto Rican children in bilingual education programs in the United States: A critique of Lloyd M. Dunn's monograph. *Hispanic Journal of Behavioral Sciences, 10*, 237–252.
- Publication Manual of the American Psychological Association (5th ed.). (2001). Washington, DC: American Psychological Association.
- Restrepo, M. A., & Silverman, S. W. (2001). Validity of the Spanish Preschool Language Scale–3 for use with bilingual children. *American Journal of Speech-Language Pathology, 10*, 382–393.
- Rogers, W. T., Gierl, M. J., Tardif, C., Lin, J., & Rinaldi, C. (2003). Differential validity and utility of successive and simultaneous approaches to the development of equivalent achievement tests in French and English. *Alberta Journal of Educational Research, 49*, 290–304.
- Rogler, L. H. (1999). Methodological sources of cultural insensitivity in mental health research. *American Psychologist, 54*, 424–433.
- Sechrest, L., Fay, T. L., & Hafeez Zaidi, S. M. (1972). Problems of translation in cross-cultural research. *Journal of Cross-Cultural Psychology, 3*, 41–56.
- Shah, C., & Long, M. (2003). Employment changes and job openings for new entrants in nursing and caring occupations in Australia. *Australian Journal of Labour Economics, 6*(3), 453–471.
- Shin, H. B., & Bruno, R. (2003, October). *Language use and English speaking ability: 2000*. Census 2000 brief. U.S. Census Bureau. Retrieved June 4, 2007, from <http://www.census.gov/prod/2003pubs/c2kbr-29.pdf>.
- Sireci, S. G., & Berberoglu, G. (2000). Using bilingual respondents to evaluate translated-adapted items. *Applied Measurement in Education, 13*, 229–248.
- Slobin, D. I. (1985). Crosslinguistic evidence for the language-making capacity. In D. I. Slobin (Ed.), *The crosslinguistic study of language acquisition: Vol. 2. Theoretical issues* (pp. 1157–1256). Hillsdale, NJ: Erlbaum.
- Standards for educational and psychological testing. (1999). Washington, DC: American Educational Research Association, American Psychological Association, National Council on Measurement in Education.
- Suizzo, M.-A. (2004). French and American mothers' childrearing beliefs: Stimulating, responding, and long-term goals. *Journal of Cross-Cultural Psychology, 35*, 606–626.
- Tamayo, J. (1987). Frequency of use as a measure of word difficulty in bilingual vocabulary test construction and translation. *Educational & Psychological Measurement, 47*, 893–902.
- Tamayo, J. (1990). A validated translation into Spanish of the WISC–R vocabulary subtest words. *Educational & Psychological Measurement, 50*, 915–921.
- Tardif, T., Gelman, S., & Xu, F. (1999). Putting the “noun bias” in context: A comparison of English and Mandarin. *Child Development, 70*, 620–635.
- Towse, J., & Saxton, M. (1998). Mathematics across national boundaries: Cultural and linguistic perspectives on numerical competence. In C. Donlan (Ed.), *The development of mathematical skills* (pp. 129–150). Hove, UK: Psychology Press/Taylor & Francis.
- Trueba, H. T. (1988). Comments on L.M. Dunn's bilingual Hispanic children on the U.S. mainland: A review of research on their cognitive, linguistic, and scholastic development. *Hispanic Journal of Behavioral Sciences, 10*, 253–262.
- Valencia, R., & Rankin, R. J. (1985). Evidence of content bias on the McCarthy scales with Mexican American children: Implications for test translation and nonbiased assessment. *Journal of Educational Psychology, 77*, 197–207.
- van der Veer, K., Ommundsen, R., Hak, T., & Larsen, K. S. (2003). Meaning shift of items in different language versions. A cross-national validation study of the Illegal Aliens Scale. *Quality & Quantity: International Journal of Methodology, 37*, 193–206.
- van der Veer, K., Ommundsen, R., Larsen, K. S., Van Le, H., Krumov, K., Pernice, R. E., et al. (2004). Structure of attitudes toward illegal immigration: Development of cross-national cumulative scales. *Psychological Reports, 94*, 897–906.
- Vigil, D. C. (2002). Cultural variations in attention regulation: A comparative analysis of British and Chinese populations. *International Journal of Language & Communication Disorders, 37*, 433–458.
- Willig, A. C. (1988). A case of blaming the victim: The Dunn monograph on bilingual Hispanic children on the U.S. Mainland. *Hispanic Journal of Behavioral Sciences, 10*, 219–236.
- Zambrano, I., & Greenfield, P. M. (2004). Ethnoepistemologies at home and at school. In R. J. Sternberg & E. L. Grigorenko (Eds.), *Culture and competence* (pp. 251–272). Washington, DC: American Psychological Association.
- Zimmerman, I., Steiner, V., & Pond, R. (1992). *Preschool Language Scale–3*. San Antonio, TX: The Psychological Corporation.
- Zimmerman, I., Steiner, V., & Pond, R. (1993). *Preschool Language Scale–3 Spanish Edition*. San Antonio, TX: The Psychological Corporation.
- Zimmerman, I., Steiner, V., & Pond, R. (2002). *Preschool Language Scale–4 Spanish Edition*. San Antonio, TX: The Psychological Corporation.
- Zuckerman, H. (1988). The sociology of science. In N. J. Smelser (Ed.), *Handbook of sociology*. (p. 511–574). Thousand Oaks: Sage.