Cross-Cultural Normative Assessment: Translation and Adaptation Issues Influencing the Normative Interpretation of Assessment Instruments

Kurt F. Geisinger

This article describes some of the issues affecting measures that are translated and/or adapted from an original language and culture to a new one. It addresses steps to ensure (a) that the test continues to measure the same psychological characteristics, (b) that the test content is the same, and (c) that the research procedures needed to document that it effectively meets this goal are available. Specifically, the notions of test validation, fairness, and norms are addressed. An argument that such adaptations may be necessary when assessing members of subpopulations in U.S. culture is proposed.

Consider the following example. A test developer has constructed a new personality inventory, standardized it, validated its primary uses in a sequential series of interrelated studies, and successfully marketed it to practicing psychologists throughout the United States. The test developer travels back and forth across the country providing lectures to psychologists, instructing them on the model underlying the instrument, and training them to use the inventory appropriately. After several years, significant research on the instrument and major national conferences discussing its use have occurred. It is widely perceived as both construct valid and clinically useful. The only negative aspect concerning the inventory is a minor controversy on the generalizability of the instrument to various underrepresented minority groups in this country. The instrument nevertheless becomes so widely used that the instrument’s publisher and some international psychologists who performed their graduate work in the United States and were trained in the use of the inventory begin to press the test author to translate the instrument into other languages.

What issues need to be considered with regard to the adaptation of an instrument to different countries, cultures, and people? If we are able to identify the problems inherent in transporting instruments to new cultures, are not some of the same issues that apply also relevant when using an instrument with various subgroups or subpopulations within a given country or population? That is, are there procedural similarities between adapting a personality inventory for use with individuals in Spain and adapting that same inventory for use with Latinos in the United States? Although the literature is replete with examples of tests translated from English into other languages, the procedures for engaging in these activities are still in their infancy, and many of the better procedures are not widely practiced.

The adaptation of assessment instruments for new target populations is generally required when the new target population differs appreciably from the original population with which the assessment device is used in terms of culture or cultural background, country, and language. Most cross-cultural adaptations of assessment instruments involve the translation of an instrument from one language into another. In some instances, however, adaptations of assessment instruments are needed even when the language remains the same, because the culture or life experiences of those speaking the same language differ. Numerous instruments have been adapted for use with groups that differ experientially from those with whom the instrument was originally used. The Minnesota Multiphasic Personality Inventory—Adolescent [MMPI-A], for example, is an adaptation of the MMPI-2 developed to be used with adolescents, whose life experiences systematically differ from those of adults. Other measures have been developed with different versions for men and women, for example, again on the basis of the generally different life experiences and resultant orientations of men and women.

Many instruments are adapted from one country, typically from the United States, given its strong orientation toward and history of success in developing many measurement instruments, to another, a so-called target country, culture, or language. In fact, using the compact disk version of the American Psychological Association’s (1974–1992) Psychological Abstracts, I found that 324 articles were cited when foreign language translation and testing were used as conjoint descriptors, whereas only an additional 36 were cited when cross-cultural and testing were used. Although there are probably multiple articles written regarding some test translations, certainly there

I would like to thank Kevin Moreland of the Psychology Department at Fordham University for providing several useful references. I would also like to thank Janet F. Carlson of the State University of New York College at Oswego for her careful reading of the article.

I found two draft documents circulated by the International Test Commission of great use in preparing this article. Because the International Test Commission does not wish the citation of specific standards at the present time from their draft document, "Standards for Adapting Instruments and Establishing Score Equivalence" (International Test Commission et al., 1993), only those principles commonly found in the literature were mentioned in this article. Any errors that remain are of course mine.

Correspondence concerning this article should be addressed to Kurt F. Geisinger, College of Arts and Sciences, State University of New York, 601 Culkin Hall, Oswego, New York 12126.
are other test translations that do not lead to any publications, only to clinical or some other applied use. Most of these translations are performed because a need exists for assessment devices in certain languages and countries or, with emerging subcultures, within a given culture (Butcher, 1985; Butcher & Pancheri, 1976).

In recent years, the term test adaptation frequently has replaced test translation. This shift in terminology documents the adaptations in references to culture, in content, and in wording that are needed in addition to simple translation in revising a test. These changes are needed to meet the requirements of a circumstance that differs qualitatively from the original use of an assessment device. Few guidelines regarding test adaptations have been developed, and those that do exist have not been widely circulated. The International Test Commission, working in conjunction with the European Association of Psychological Assessment, the International Association of Applied Psychology, the International Association of Cross-Cultural Psychology, the International Association for the Evaluation of Educational Achievement, the International Language Testing Association, and the International Union of Psychological Science, has begun systematizing the procedures that are recommended in test adaptation.

The technical literature for guiding test translations, or test adaptations, as they are sometimes called, is rather incomplete (from a measurement perspective), and scattered through a plethora of international journals, reports and books. Also, the more complex measurement methods (e.g., item response models and factor analysis) which appear to be useful in formally establishing the equivalence of scores obtained from tests translated (adapted) into multiple languages are not well-known by persons involved in test translations research. (International Test Commission, 1993, p. 1)

The 1985 edition of the Standards for Educational and Psychological Testing (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, 1985) added a chapter on the testing of individuals who are linguistic minorities, a chapter not included in the previous versions. A number of the seven standards provided in that chapter are also applicable to the problems involved in cross-cultural measurement.

The present article addresses four questions. First, how should a measure from one language and one culture be adapted to another language and culture? Second, how does one know whether a measure adapted to a new language and/or culture measures the same construct or constructs that it did in the first? A third question relates closely to the second: Is the newly adapted measure useful once it has been “fitted” to a new culture and language? Fourth, do scores from the new instrument mean the same thing that they do in the initial culture and language?

Procedures for Adapting Assessment Measures for New Populations

Anytime a measure is simply used with a population that differs qualitatively from the one for which it was originally developed, one must check its continued validity and usefulness in that new population, even if the test itself remains unchanged. When it is likely that a measure needs to be adapted for use with a new population, the onus is even greater on those adapting the measure to demonstrate its usefulness.

Does a Given Measure Need to Be Adapted?

An early decision to be made is whether or not an assessment device actually needs to be adapted for a new intended use. The answer to this question is generally easy when applied to an extreme circumstance. When a new target population is not likely to differ appreciably from the original population with whom the assessment device was used, it is unlikely that the measure needs to be adjusted to that new population. A measure that has been developed with students from Pennsylvania State University would not likely need to be adapted if it were to be administered to students from the University of North Carolina, for example. (However, if it were found that some items on the instrument used local vernacular or were based on information known only to residents of central Pennsylvania, then some minor accommodations would be needed.) At the other extreme, when one wishes to administer a measure to individuals who cannot communicate in the language in which the measure was originally written, then the measure obviously must be translated and perhaps otherwise altered. Indeed, such a translation or adaptation must typically consider cultural as well as language differences between the original and the target populations.

The two situations that lead to more difficult test adaptation decisions relate to subpopulations within a given nation and cultural adaptations within a single language. Whether measures built in the United States primarily with White, middle-class Americans need to be adapted for use with African Americans, Asian Americans, Latinos or Hispanic Americans, and Native Americans, for example, can be a difficult question, sometimes with complex answers. Pennock-Roman (1990, 1992), for example, analyzed the use of the Scholastic Aptitude Test (SAT) in making college admissions decisions with Hispanic American applicants. She found that, for some individuals, notably those with a high degree of English language ability, the validity of the use of the English-language version of the SAT indicated that the use of the test in this manner was appropriate. For others, however, she found that the validity differed, and its use with these individuals was less appropriate. Factors such as quality of educational background, parents’ socioeconomic status, familiarity with tests, and specific test-taking skills and abilities also influenced the appropriateness of such test use. Thus, in some cases (e.g., with first-generation Latinos and some Asian Americans and Native Americans), language issues may be of major consequence. In others, cultural issues are paramount and require that a measure be adapted before it is used. Descriptions of adaptations of the MMPI for Turkey, Hong Kong, Greece, and Chile may be found in Savasir and Erol (1990), Cheung (1985), Manos (1985), and Rissetti and Malles (1985), respectively, and serve as excellent examples of both language and cultural adaptations of that particular measure.

Steps for Adapting a Measure

The procedures provided later are intended to serve as general guidelines for adapting an assessment instrument to a new
The use of these procedures in cross-cultural testing situations may be found in O'Brien, 1992, or van de Vijver & Poortinga, 1991. For a brief discussion of some of the techniques themselves, see Cole & Moss, 1989.) Many of the differential item functioning approaches compare the difficulty of specific items across two groups after controlling for overall ability level. Overall ability is generally controlled by blocking members of both groups studied by an overall indication of the characteristic in question. Total test score, or some
derivative thereof, is most commonly used as the indicator of ability. One statistical test that can be used to evaluate whether an item is biased, then, is the chi-square test. Such an analysis might be applied in the present example with the following three factors: (a) group membership (a sample that took the instrument in question in the original language and one that took the translated version of the instrument); (b) level of score on the entire instrument (e.g., low, average, and high); and (c) response (e.g., correct versus incorrect). If the chi-square value is statistically significant, it would indicate that the item is differentially more difficult in one language than in the other, at least at some score levels.

To understand the use of this approach, imagine an item on an intelligence scale and its translated version. Suppose that the item is differentially more difficult in the second language (and, hence, sometimes considered as biased). Differential item functioning can identify that the item is more difficult in the second language, but we would still not know why the item is more difficult when adapted. Is it because it is less intelligible than the first, because the sample of the second group is not as representative as it should have been, or some other reason? Until these questions can be answered definitively, one must strongly consider revising the tentatively translated or adapted item.

6. Standardize the scores. If desirable and appropriate, equate them with scores on the original version. Material concerned with these procedures is presented in response to the fourth question addressed by the present article. At this time, it is sufficient to state that most instruments that are translated or adapted for use with a new population, unless they are used purely as research measures, especially those research measures that are unpublished, use a standard scale such as the T scale (mean of 50, standard deviation of 10) to present their scores. Using data from the field testing, if the sample is large enough (probably at least 750-1,000 participants, depending on how many subgroups in the population need adequate representation), it probably is possible to perform a standardization. If the sample is not large enough or fully representative, then collection of a new standardization sample is needed. The analytic procedures involved in such processes are straightforward and provided in Peterson, Kolen, and Hoover (1989). It must be noted, however, that it would rarely, if ever, be appropriate to use the norms from the original instrument as calculated on the original population with the revised and adapted instrument. That is, one cannot simply translate a measure into a second language and use the norm tables for that measure based on the first language version of the instrument. There are several reasons why using the original language norms would not be appropriate. First, because the instrument has not yet been validated in the new language, it is not even clear that the instrument measures the same characteristic as the instrument in the original language. Second, assuming that the instrument continues to be valid and that the populations in both cultures are comparable, some of the questions on the instrument may have changed enough in meaning to affect the pattern of responses. Hence, the scores resulting from the newly translated instrument may be elevated, reduced, or changed in some other way. Third, assuming that the instrument continues to measure the same characteristic and the underlying characteristic has a comparable meaning in the second culture, it is possible that the two cultural groups differ systematically in the underlying construct. "Cross-cultural researchers need to be aware that some variables from highly different cultures differ in form and quantity across groups" (Butcher, in press-c, p. 6). Many cultures, for example, differ with regard to their need for achievement, need for affiliation, time orientation, and so on. Butcher and Han (in press) provide a well-developed overview of some of the methods for establishing the cross-cultural equivalence of measures that have been translated and otherwise adapted for a new cultural or linguistic population.

Substantial evidence of the comparability both of group performances on the two measures and of validation would be needed before such use could possibly be appropriate. Similarly, the interpretative judgments learned by experienced clinicians over years of practice cannot simply be applied to the new version for the same reasons. However, one might have relatively more faith in using the U.S. norms for a translated measure if the test has already been translated into a number of other languages and it has been found that U.S. norms well approximate the new cultural norms (Butcher, in press-c). Such information might be seen as the cross-cultural generalizability of norms. Butcher and Garcia (1978) have found that U.S. norms can frequently be used successfully in other countries. However, engaging in such activity would appear to contradict standards 4.3 and 13.6 of the Standards for Educational and Psychological Testing (American Educational Research Association et al., 1985), unless a strong justification for such a use could be provided. These two standards call for (a) the use of norms developed for clearly described populations and, hence, the frequent use of locally developed norms and (b) evidence of test comparability when tests are translated into a different language, respectively. A strong rationale, preferably coupled with empirical evidence, is required, therefore, before U.S. norms are used with culturally and linguistically different populations.

A potential user may learn of a translated version of an instrument and seek to learn whether there are norms (and validation evidence) for the translated version. To learn such information, such individuals should contact the test publisher in the second culture, if there is one. Alternatively, a user may contact the original test publisher from the original country, the translator, and any individual who may have used the translated instrument, whether for practical work or research. In addition to published research, theses and dissertations using the instrument may be a helpful source of normative data bearing on both the usefulness and the scoring of the translated instrument, even if these purposes are not among the explicit intents of the research study. Such studies often provide means and standard deviations, if not the entire distribution of scores resulting from the instrument. A potential test user could perhaps begin by using an instrument in conjunction with such data, if the population with which the research was performed were appropriate and similar to that with which the instrument would be used professionally.

7. Perform validation research as appropriate. The next section of this article presents information on the research needed to document that (a) an assessment device measures the same qualities in both languages or versions and (b) the new version continues to provide scores that are interpretable in the manner proposed. Standard 6.2 of the Standards for Educational and
to complete the assessment device need to be clear and simple. A Few Cautions

be translated from one language into another, a number of cau-
sations across cultures before using the test, and has indeed sug-

Such research is almost always necessary.

8. Develop a manual and other documents for the users of the assessment device. “Test developers have a responsibility to pro-

ten so as to minimize potential misunderstandings. One should

standard 6.6 of the Standards for Educational and Psychological Testing (American Educational Research As-

To ensure that users, even professional users, are informed about a translated or revised assessment device, it would be ex-

10. Collect reactions from users. As one implements the newly revised instrument, it is frequently propitious to gather com-

A Few Cautions

If an assessment device is adapted for use with individuals in a new culture, and especially if the assessment device needs to be translated from one language into another, a number of cau-

mat for the culture. When different cultural or national groups vary in their levels of sophistication with differing item formats, a sufficient number of exemplary practice exercises should be used. Incidents and situations depicted in items in both the original and target languages should be equally common in their occurrence as well as similar in behavioral and construct interpretation. Vocabulary should be similar in both versions. Text (e.g., test questions), as stated previously, should not simply be translated from one language into the other. When straight translations are used in the development of assessment devices, the words used often differ in the frequency of their use, their difficulty for or familiarity to members of the culture, and their connotations.

Does the Measure Assess the Same Constructs in the New Language or Culture?

Standard 13.4 of the Standards for Educational and Psychological Testing (American Educational Research Association et al., 1985) states that, “when a test is translated from one language or dialect to another, its reliability and validity for the uses intended in the linguistic groups to be tested should be established” (p. 75). Two changes in test use necessitate the reestablishment of reliability and validity when a measure is adapted for use with those who differ from the original group for which the measure was intended either culturally or linguistically. First, any time a test is altered, especially if it is changed appreciably, its reliability and validity need to be checked; it is possible that the alterations will affect the validity of the measure for the intended use. Second, if a test is applied to a new population, the test adapter needs to demonstrate that the instrument remains to assess the same qualities with the same degree of accuracy in the new population. This necessity is especially pertinent when the new population differs from the original population in terms of cultural development.

That a measure needs to be studied to determine whether it measures the same constructs after it has been adapted to a new linguistic group is a matter of validation in general and construct validation in particular (Geisinger, 1992b). The Standards for Educational and Psychological Testing (American Educational Research Association et al., 1985) suggest that validation information may come from three sources: construct-related evidence, content-related validation evidence, and criterion-related evidence. The latter two approaches are discussed later. Construct-related validation depends on a wide variety of research designs, techniques, and types of information. (For a complete description of the types of information that may be considered as part of construct validation, see Geisinger, 1992b, or Messick, 1989.)

One common source of construct-related validation information is the factor-analytic technique applied to test data. Factor-analytic techniques probably have been the most frequently used procedures in the evaluation of tests adapted to linguistically different populations (e.g., Cattell, 1970). Cattell has “always been fully aware of the necessity of comparing factor structures across cultures before using the test, and has indeed suggested a rigorous and original method of carrying out such comparisons” (Eysenck & Eysenck, 1983, p. 43). Eysenck and Eysenck (1983) suggested that determining that a translated or
otherwise adapted measure covers the same dimensions and covers them in the same quantities in a linguistically different population is critical to our interpretation and use of scores that result from the measure. (Gorsuch, 1983, has provided a full description of some of the factor-analytic techniques that can be used to make these comparisons.) In some cases, it has been found that adapted measures do assess the same dimensions across linguistic groups, but that the loadings of items on factors differ across cultures. Such findings indicate that the dimensions are indeed found across nations, but that the dimensions differ somewhat cross-culturally. For example, Eysenck and Eysenck reported that,

in most cases a few items change loadings drastically, ... Some of these losses and shifts could be explained by inadequate translation; ... factor analysis is capable of uncovering such errors. ... More interesting are changes in loading due to cultural differences, and we have usually found that it is possible to account for these observed changes in terms of fairly obvious suggestions as to causation made by representatives of the country in question, who were involved in the translation and/or administration of the test. (Eysenck & Eysenck, 1983, pp. 46, 48)

Interpretation of personality or other assessed variables is dependent on the finding that items adapted from those correlating together highly and apparently measuring the same construct in the original language continue to form a similarly interpretable factor in the target language and with data from the assessments made of individuals from the new culture and in the new language. An excellent example of a cross-cultural, factor-analytic investigation can be found in Irvine and Carroll (1980).

Ben-Porath (1990) compared cross-cultural factor-analytic evidence for five personality inventories. He argued against using the traditional exploratory factor-analytic techniques that frequently have been used in both the study of personality and, more specifically, the cross-cultural study of personality assessment. Many confirmatory factor-analytic models have become available with structural equation modeling and maximum-likelihood factor-analytic techniques. Such techniques permit one to perform a hypothesis test confirming that an adapted assessment device measures the same qualities, in the same proportions, and with the same factor structures (i.e., item–factor loadings), as does the original measure. Ben-Porath suggested that the number of factors in the target population be restricted to that found in the original population.

Samples needed for cross-cultural, factor-analytic comparisons must be large and representative of the population from which they are drawn. When one is making factor-analytic comparisons across cultural groups, to perform accurate contrasts, one needs to control for sex, age, and other relevant variables. Such controls might be performed by equalizing, or nearly equalizing, the numbers of individuals who are male and female and who fall within each age level in the cross-cultural groups. Another such variable to consider controlling might be educational level, especially when one is dealing with measures of cognitive and educational performance. Clearly, however, a variable such as educational level would be rather difficult (and perhaps impossible) to control when adapting a measure from a developed, Western nation for use in a developing nation. That is, the educational levels may vary so radically between these two cultures that, if the two samples were forced to be equated, the resultant samples would likely have little external validity (Cook & Campbell, 1979).

It is also incumbent on those performing comparative factor-analytic investigations to review distributions for range restrictions and outliers (Ben-Porath, 1990) and to evaluate the data as one normally does with complex data with multicolinearity.

A question needing consideration is whether it is reasonable to compare assessment instruments in terms of their factor structures when the original measures were themselves neither developed using factor-analytic methods nor justified for use on the basis of factor-analytic results. The MMPI, for example, has been found to yield a reasonably stable factor structure across a variety of cultures (Ben-Porath, 1990). However, the use of the MMPI has been generally justified on the basis of its criterion-related validation, not on the basis either of its construct validation in general or its factor structure in particular. Other instruments that have been justified on the basis of their factor structures and construct validity (e.g., the Eysenck Personality Questionnaire and the Sixteen Personality Factor Questionnaire) have provided more mixed cross-cultural, factor-analytic evidence. It would be more important that the author of a measure, the original development and use of which were heavily dependent on factor analysis, provide a similar factor structure when adapted for a new language or culture. Butcher and Han (in press) reviewed factor-analytic techniques for documenting that a translated measure is equivalent to the original version. They called special attention to the distinction between confirmatory and exploratory factor-analytic techniques; the former techniques provide a significant improvement over many of the exploratory factor-analytic techniques, the results from which must be evaluated from a highly subjective perspective.

The factor approach procedures are the primary construct-validation technique that has been used in the literature. It would be helpful if other construct validation designs (e.g., the multimethod–multitrait matrix or responsiveness to experimental manipulations) were used as well.

Content-related validation and criterion-related validation have been frequently used to justify the use of translated measures in individual studies. Cross-cultural, content-related validation studies are difficult to perform. Content-related validation is frequently used to justify measures of educational achievement. Imagine a measure that has been carefully constructed to assess a given school subject matter in a representative fashion. Is it likely that such a measure would be equally representative in a different country and a different language, where schools and curricula differ markedly from that in the original population? Such a situation might only be sensible when individuals speaking two different languages attend the same school and learn the same material from the same curriculum in different languages, but what is the likelihood that another culture or country would have the same exact domain of content?

Similarly, the criteria used in criterion-related validation studies are likely to differ substantially across cultures. School success is defined differently even among different Western nations. Even some seemingly concrete criteria such as hospitalization (used as a criterion against which a psychodiagnostic instrument might be evaluated) will fluctuate in meaning across
cultures because cultural definitions of mental illness and the base rates for identifying such illness differ markedly. For these reasons, Ben-Porath (1990) has stated that,

if in developing measuring instruments in the field of personality we were interested only in predictive or concurrent validity, we would not be obliged to study those instruments in other cultures. Each culture could conceivably develop its own optimal empirically keyed instrument. If, on the other hand, we wish to understand psychologically what a given instrument measures, that is, if we are interested in construct validity, then we may well want to broaden and deepen this understanding by finding out whether meanings and interpretations associated with an inventory can accompany its use in another [culture], (p. 28)

Hence, for the reasons cited, construct-related validation likely will continue to prove most useful in evaluating measures that are used cross-culturally.

Is the Measure Useful in the New Culture?

Typically, assessment devices are adapted for new settings because a need for that measure is apparent in the target culture. Furthermore, the developers of most well-constructed measures have taken theoretical and validity concerns into account in the developmental process. The decision to adapt an existing measure from one culture to another implies that the translator or adapter believes that the same constructs will lead to the same or similar interpretations and predictions in the new culture. Thus, those adapting a measure may not always need to address all of the validity concerns present in the construction of a truly new measure. The mapping of the components of a construct may be similar across cultures, for example. A determination of whether a measure that has been successfully adapted to a new country, culture, or language is useful in that new setting is largely a matter of validation, often criterion-related validation, although we know that many factors influence the usefulness of a measure in addition to validity (e.g., base rates, costs, and so forth). Therefore, although criterion-related validation may not prove useful in answering questions of interest to cross-cultural researchers, it will continue to serve as a primary technique for evaluating the usefulness of a measure within a given culture.

What Do Scores on the Adapted Measure Mean?

The interpretation of any score on an assessment device is dependent on many kinds of information. Clearly, validation (including information on fairness and bias; see Geisinger, 1992a) and reliability information are critical to test interpretation. Normative information carries meaning to a professional regarding the placement of an individual within a population distribution of test takers, permitting the professional to interpret the likely meaning of the score. This interpretative meaning of test scores is especially useful for tests with well-established validation information and norms. Thus, a professional admissions counselor at a university can look at an individual's scores on the SAT or a similar measure and approximate the probability that the prospective student will succeed at the university. Similarly, an informed admissions committee making decisions with regard to the applications for graduate study in psychology can consider their applicants' Graduate Record Examination® scores by using their accumulated (validation) information on how former applicants have done in graduate study and both their local norms (whether actually computed or not) and the examination's national test norms. Clinicians use scores from well-established measures such as the MMPI in much the same way. Experienced test users can make highly skilled interpretations on the basis of the proper test information, especially if they are also able to gather the other information that they need to fine-tune their judgments.

Just translating a test and using the same scoring algorithm may not lead to meaningful scores, even when the understanding of specific scores is well-established in the original language, nation, or culture. Cultural and other differences between the original and target populations as well as language differences across the two forms of the assessment device may render the use of the same scores to denote identical meaning across test forms and cultures substantially less meaningful. For example, research using the MMPI with various Chinese samples has led to the observation of elevated scores on several scales . . . among Chinese college students in the U.S. and in Singapore using the original English version of the MMPI . . . . Results on a 550 item Chinese version of the Group Form translated in Taiwan suggested the necessity to rescale the inventory so that the meaning would fit Chinese cultural patterns. (Lu, 1967, as quoted by Cheung, 1985, p. 131)

In general, almost all tests that are adapted to a new language or culture will need to be renormed so that scores on the new version of the assessment device may be studied. Moreover, validation evidence needs to be gathered to document the meaning of scores on the new version. Scores that are "numerically identical can have a psychologically different meaning across cultures" (van de Vijver & Poortinga, 1991, p. 279). "Considerable research has also demonstrated that culture influences the rates, manifestations, course, and outcomes of mental disorders" (Butcher, in press-c, p. 13). Before interpretation can be meaningfully applied to test results, evidence of both construct and instrument comparability across cultures is needed.

O'Brien (1992) provided an example of the concomitant construction, norming, and equating of two forms of a language proficiency examination for Anglos and Hispanics using an elementary item-response theory approach (see also O'Brien, 1989). This chapter would make useful reading for anyone interested in either simultaneous test construction in two or more languages or adapting an assessment device for a new culture (a much more common procedure). His procedures can be applied to studying the use of a single measure for two populations or the development of two different measures, one for each language group. The first four of the five steps in scale construction that O'Brien advocated (scale development, item calibration, item validation, and person–measure validation) would typically be performed in any instrument development process based on item response theory. The fifth step, however, relates to population congruence.

1 Graduate Record Examination is a registered trademark of Educational Testing Service.
This step requires two procedures. First, items calibrated separately in Anglo and Hispanic samples are plotted against an identity line in a bivariate graph and bounded by 95% confidence bands. Items falling within the bands are invariant and evidence the same underlying calibration of difficulty. Items outside the bands shift position and have different meanings across the two samples. For purposes of comparing Anglo and Hispanic samples, only invariant items should be used for person measurement. (O'Brien, 1992, p. 48)

O'Brien thus recommended using only those items of comparable difficulty levels in each of the two forms. (This procedure is comparable to the analysis of differential item functioning presented earlier in this article.) He developed two parallel forms (English and Spanish) of this language proficiency instrument, the so-called Language Assessment Battery (LAB). Both forms were vertically equated so that scaled scores could be compared across the two versions of the examination,

on the assumption that English and Spanish language proficiency are directly comparable. Further research is necessary to determine the criteria on which this assumption may be validated for a given test. In the case of LAB, this assumption seemed reasonable, given that items were paired by content and difficulty across English and Spanish. (O'Brien, 1992, pp. 48–49)

Such a scoring procedure permits students' scores with respect to their English and Spanish proficiencies, for example, to be directly compared, an obvious advantage. For those translators of assessment devices seeking comparable interpretations of scores on both language versions, such a procedure would be highly recommended. Without evidence of external validation, however, such information is a necessary but insufficient requirement for comparable score interpretation. Methods for equating such test forms may be found in Angoff (1971) and Peterson, Kolen, and Hoover (1989).

Summary

The translation and adaptation of psychological measures from English to other languages is currently a widespread activity. These efforts require both language and cultural sensitivities; in past years, some measures were only translated linguistically. Little care was taken to adapt measures for cultural differences. Such steps must be taken, and the issue of cultural adaptation may make it difficult to directly translate and use items from some measures and, in some cases, to translate and use items from entire measures. The steps recommended for adapting measures will help ensure the proper use of adapted measures.

Although psychometric theory has lagged in this area, psychometric procedures are available to study adapted measures to attempt to determine if they are assessing the same constructs in the adapted form, if the scores may be used with equivalent validity, if bias either on the item level or with respect to the entire instrument is present, and if scores are equally meaningful on both forms. Pilot testing and subsequent large-scale field testing of adapted measures are required by professional standards and practice. Knowledge that a revised version of an instrument measures what is intended and what it measures in the original language or cultural version is essential information that must be gathered before adapted measures might be used in practice. Those who adopt measures also need to be sensitive to the training and documentation needs of the users of the assessment instrument. If such developmental practices are followed, the use of the measures that result are most likely to yield beneficial results. When steps are bypassed in the construction, translation, and validation of translated measures, responsible test use is less likely to be possible.

References


Eysenck, H. J., & Eysenck, S. B. G. (1983). Recent advances in the


Received January 5, 1994
Revision received May 16, 1994
Accepted May 18, 1994

---

**New Editors Appointed, 1996–2001**

The Publications and Communications Board of the American Psychological Association announces the appointment of three new editors for 6-year terms beginning in 1996. As of January 1, 1995, manuscripts should be directed as follows:

- For *Behavioral Neuroscience*, submit manuscripts to Michela Gallagher, PhD, Department of Psychology, Davie Hall, CB# 3270, University of North Carolina, Chapel Hill, NC 27599.

- For the *Journal of Experimental Psychology: General*, submit manuscripts to Nora S. Newcombe, PhD, Department of Psychology, Temple University, 565 Weiss Hall, Philadelphia, PA 19122.

- For the *Journal of Experimental Psychology: Learning, Memory, and Cognition*, submit manuscripts to James H. Neely, PhD, Editor, Department of Psychology, State University of New York at Albany, 1400 Washington Avenue, Albany, NY 12222.

Manuscript submission patterns make the precise date of completion of 1995 volumes uncertain. The current editors, Larry R. Squire, PhD, Earl Hunt, PhD, and Keith Rayner, PhD, respectively, will receive and consider manuscripts until December 31, 1994. Should either volume be completed before that date, manuscripts will be redirected to the new editors for consideration in 1996 volumes.